

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](#). 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](#). 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](#). 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:

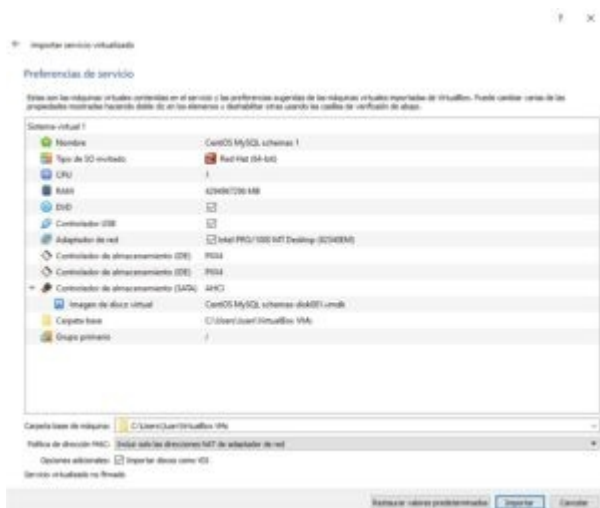
Pulsamos, Archivo, Importar servicio virtualizado y buscamos el archivo:

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

```
File Edit View Search Terminal Help
[cloudera@localhost ~]$ mysql -ucloudera -pCl@7d3r4
mysql: [Warning] Using a password on the command line interface can be insecure.
Welcome to the MySQL monitor.  Commands end with ; or \g.
Your MySQL connection id is 9
Server version: 8.0.31 MySQL Community Server - GPL
Copyright (c) 2000, 2022, Oracle and/or its affiliates.
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.

mysql> show databases;
+-----+
| Database |
+-----+
| Cloudera |
| books    |
| demobid  |
| information schema |
| mysql    |
| performance schema |
| retail_db |
| sys      |
+-----+
8 rows in set (0.03 sec)

mysql> use demobid;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
mysql> show tables;
+-----+
| Tables in demobid |
+-----+
| books              |
| countries           |
| departments         |
| dept               |
| emp                |
| emp_details_view   |
| employees            |
| job_history         |
+-----+
```

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 demobid;

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;

```
mysql> select * from departments limit 10;
+-----+-----+-----+-----+
| department_id | department_name | manager_id | location_id |
+-----+-----+-----+-----+
| 10 | Administration | 200 | 1700 |
| 20 | Marketing | 201 | 1800 |
| 30 | Purchasing | 114 | 1700 |
| 40 | Human Resources | 203 | 2400 |
| 50 | Shipping | 121 | 1500 |
| 60 | IT | 103 | 1400 |
| 70 | Public Relations | 204 | 2700 |
| 80 | Sales | 145 | 2500 |
| 90 | Executive | 100 | 1700 |
| 100 | Finance | 108 | 1700 |
+-----+-----+-----+-----+
10 rows in set (0.01 sec)
```

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](#).

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.

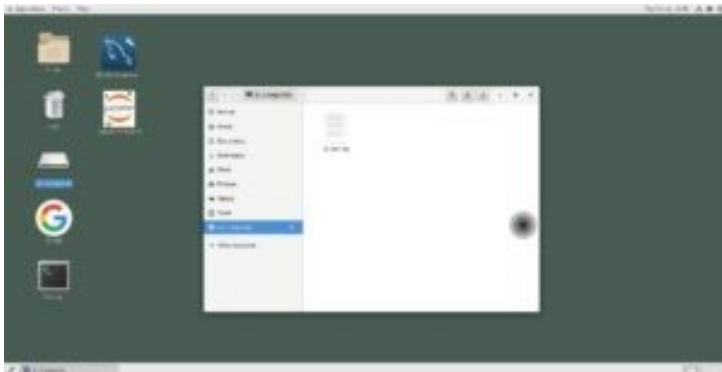
Veamos un ejemplo de uso de estos ficheros de script de consultas SQL. Para ello vamos a ver en primer lugar como nos podemos comunicar, como podemos transferir ficheros entre la máquina virtual y la máquina anfitriona (Windows). Bien, pues esta es otra de las ventajas de trabajar con una máquina virtual 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:

```
Code  +  [icon]  *script1.sql
of Computer (Virtual) Compartida

system clear
use retail db;
DROP VIEW IF EXISTS estudio_orders1;
DROP VIEW IF EXISTS estudio_orders2;
DROP VIEW IF EXISTS estudio_orders3;
DROP VIEW IF EXISTS estudio_orders4;
create view estudio_orders1 as select order_item.product_id, products.product_name,
round(sum(order_item.subtotal)) as Total, round(sum(order_item.subtotal)) as Minimo,
round(max(order_item.subtotal)) as Maximo, round(sum(order_item.subtotal)) as Medio from order_items
join products on order_items.order_item.product_id=products.product_id group by order_item.product_id;
create view estudio_orders2 as select round(sum(total)) as Total, round(max(total)) as Maximo,
round(min(total)) as Minimo, round(avg(total)) as Medio from estudio_orders1;
select * from estudio_orders2;
create view estudio_orders3 as select if(total=0 AND total<10000,1,0) as caso1, if(total=10000
AND total<20000,1,0) as caso2, if(total=20000 AND total<30000,1,0) as caso3, if(total=30000
AND total<40000,1,0) as caso4, if(total=40000 AND total<50000,1,0) as caso5, if(total=50000
AND total<60000,1,0) as caso6, if(total=60000 AND total<70000,1,0) as caso7, if(total=70000
AND total<80000,1,0) as caso8, if(total=80000 AND total<90000,1,0) as caso9, if(total=90000
AND total<100000,1,0) as caso10, if(total=100000 AND total<110000,1,0) as caso11, if(total=110000
AND total<120000,1,0) as caso12, if(total=120000 AND total<130000,1,0) as caso13, if(total=130000
AND total<140000,1,0) as caso14, if(total=140000 AND total<150000,1,0) as caso15, if(total=150000
AND total<160000,1,0) as caso16, if(total=160000 AND total<170000,1,0) as caso17, if(total=170000
AND total<180000,1,0) as caso18, if(total=180000 AND total<190000,1,0) as caso19, if(total=190000
AND total<200000,1,0) as caso20, if(total=200000 AND total<210000,1,0) as caso21, if(total=210000
AND total<220000,1,0) as caso22, if(total=220000 AND total<230000,1,0) as caso23, if(total=230000
AND total<240000,1,0) as caso24, if(total=240000 AND total<250000,1,0) as caso25, if(total=250000
AND total<260000,1,0) as caso26, if(total=260000 AND total<270000,1,0) as caso27, if(total=270000
AND total<280000,1,0) as caso28, if(total=280000 AND total<290000,1,0) as caso29, if(total=290000
AND total<300000,1,0) as caso30, if(total=300000 AND total<310000,1,0) as caso31, if(total=310000
AND total<320000,1,0) as caso32, if(total=320000 AND total<330000,1,0) as caso33, if(total=330000
AND total<340000,1,0) as caso34, if(total=340000 AND total<350000,1,0) as caso35, if(total=350000
AND total<360000,1,0) as caso36, if(total=360000 AND total<370000,1,0) as caso37, if(total=370000
AND total<380000,1,0) as caso38, if(total=380000 AND total<390000,1,0) as caso39, if(total=390000
AND total<400000,1,0) as caso40, if(total=400000 AND total<410000,1,0) as caso41, if(total=410000
AND total<420000,1,0) as caso42, if(total=420000 AND total<430000,1,0) as caso43, if(total=430000
AND total<440000,1,0) as caso44, if(total=440000 AND total<450000,1,0) as caso45, if(total=450000
AND total<460000,1,0) as caso46, if(total=460000 AND total<470000,1,0) as caso47, if(total=470000
AND total<480000,1,0) as caso48, if(total=480000 AND total<490000,1,0) as caso49, if(total=490000
AND total<500000,1,0) as caso50, if(total=500000 AND total<510000,1,0) as caso51, if(total=510000
AND total<520000,1,0) as caso52, if(total=520000 AND total<530000,1,0) as caso53, if(total=530000
AND total<540000,1,0) as caso54, if(total=540000 AND total<550000,1,0) as caso55, if(total=550000
AND total<560000,1,0) as caso56, if(total=560000 AND total<570000,1,0) as caso57, if(total=570000
AND total<580000,1,0) as caso58, if(total=580000 AND total<590000,1,0) as caso59, if(total=590000
AND total<600000,1,0) as caso60, if(total=600000 AND total<610000,1,0) as caso61, if(total=610000
AND total<620000,1,0) as caso62, if(total=620000 AND total<630000,1,0) as caso63, if(total=630000
AND total<640000,1,0) as caso64, if(total=640000 AND total<650000,1,0) as caso65, if(total=650000
AND total<660000,1,0) as caso66, if(total=660000 AND total<670000,1,0) as caso67, if(total=670000
AND total<680000,1,0) as caso68, if(total=680000 AND total<690000,1,0) as caso69, if(total=690000
AND total<700000,1,0) as caso70, if(total=700000 AND total<710000,1,0) as caso71, if(total=710000
AND total<720000,1,0) as caso72, if(total=720000 AND total<730000,1,0) as caso73, if(total=730000
AND total<740000,1,0) as caso74, if(total=740000 AND total<750000,1,0) as caso75, if(total=750000
AND total<760000,1,0) as caso76, if(total=760000 AND total<770000,1,0) as caso77, if(total=770000
AND total<780000,1,0) as caso78, if(total=780000 AND total<790000,1,0) as caso79, if(total=790000
AND total<800000,1,0) as caso80, if(total=800000 AND total<810000,1,0) as caso81, if(total=810000
AND total<820000,1,0) as caso82, if(total=820000 AND total<830000,1,0) as caso83, if(total=830000
AND total<840000,1,0) as caso84, if(total=840000 AND total<850000,1,0) as caso85, if(total=850000
AND total<860000,1,0) as caso86, if(total=860000 AND total<870000,1,0) as caso87, if(total=870000
AND total<880000,1,0) as caso88, if(total=880000 AND total<890000,1,0) as caso89, if(total=890000
AND total<900000,1,0) as caso90, if(total=900000 AND total<910000,1,0) as caso91, if(total=910000
AND total<920000,1,0) as caso92, if(total=920000 AND total<930000,1,0) as caso93, if(total=930000
AND total<940000,1,0) as caso94, if(total=940000 AND total<950000,1,0) as caso95, if(total=950000
AND total<960000,1,0) as caso96, if(total=960000 AND total<970000,1,0) as caso97, if(total=970000
AND total<980000,1,0) as caso98, if(total=980000 AND total<990000,1,0) as caso99, if(total=990000
AND total<1000000,1,0) as caso100, if(total=1000000 AND total<1010000,1,0) as caso101, if(total=1010000
AND total<1020000,1,0) as caso102, if(total=1020000 AND total<1030000,1,0) as caso103, if(total=1030000
AND total<1040000,1,0) as caso104, if(total=1040000 AND total<1050000,1,0) as caso105, if(total=1050000
AND total<1060000,1,0) as caso106, if(total=1060000 AND total<1070000,1,0) as caso107, if(total=1070000
AND total<1080000,1,0) as caso108, if(total=1080000 AND total<1090000,1,0) as caso109, if(total=1090000
AND total<1100000,1,0) as caso110, if(total=1100000 AND total<1110000,1,0) as caso111, if(total=1110000
AND total<1120000,1,0) as caso112, if(total=1120000 AND total<1130000,1,0) as caso113, if(total=1130000
AND total<1140000,1,0) as caso114, if(total=1140000 AND total<1150000,1,0) as caso115, if(total=1150000
AND total<1160000,1,0) as caso116, if(total=1160000 AND total<1170000,1,0) as caso117, if(total=1170000
AND total<1180000,1,0) as caso118, if(total=1180000 AND total<1190000,1,0) as caso119, if(total=1190000
AND total<1200000,1,0) as caso120, if(total=1200000 AND total<1210000,1,0) as caso121, if(total=1210000
AND total<1220000,1,0) as caso122, if(total=1220000 AND total<1230000,1,0) as caso123, if(total=1230000
AND total<1240000,1,0) as caso124, if(total=1240000 AND total<1250000,1,0) as caso125, if(total=1250000
AND total<1260000,1,0) as caso126, if(total=1260000 AND total<1270000,1,0) as caso127, if(total=1270000
AND total<1280000,1,0) as caso128, if(total=1280000 AND total<1290000,1,0) as caso129, if(total=1290000
AND total<1300000,1,0) as caso130, if(total=1300000 AND total<1310000,1,0) as caso131, if(total=1310000
AND total<1320000,1,0) as caso132, if(total=1320000 AND total<1330000,1,0) as caso133, if(total=1330000
AND total<1340000,1,0) as caso134, if(total=1340000 AND total<1350000,1,0) as caso135, if(total=1350000
AND total<1360000,1,0) as caso136, if(total=1360000 AND total<1370000,1,0) as caso137, if(total=1370000
AND total<1380000,1,0) as caso138, if(total=1380000 AND total<1390000,1,0) as caso139, if(total=1390000
AND total<1400000,1,0) as caso140, if(total=1400000 AND total<1410000,1,0) as caso141, if(total=1410000
AND total<1420000,1,0) as caso142, if(total=1420000 AND total<1430000,1,0) as caso143, if(total=1430000
AND total<1440000,1,0) as caso144, if(total=1440000 AND total<1450000,1,0) as caso145, if(total=1450000
AND total<1460000,1,0) as caso146, if(total=1460000 AND total<1470000,1,0) as caso147, if(total=1470000
AND total<1480000,1,0) as caso148, if(total=1480000 AND total<1490000,1,0) as caso149, if(total=1490000
AND total<1500000,1,0) as caso150, if(total=1500000 AND total<1510000,1,0) as caso151, if(total=1510000
AND total<1520000,1,0) as caso152, if(total=1520000 AND total<1530000,1,0) as caso153, if(total=1530000
AND total<1540000,1,0) as caso154, if(total=1540000 AND total<1550000,1,0) as caso155, if(total=1550000
AND total<1560000,1,0) as caso156, if(total=1560000 AND total<1570000,1,0) as caso157, if(total=1570000
AND total<1580000,1,0) as caso158, if(total=1580000 AND total<1590000,1,0) as caso159, if(total=1590000
AND total<1600000,1,0) as caso160, if(total=1600000 AND total<1610000,1,0) as caso161, if(total=1610000
AND total<1620000,1,0) as caso162, if(total=1620000 AND total<1630000,1,0) as caso163, if(total=1630000
AND total<1640000,1,0) as caso164, if(total=1640000 AND total<1650000,1,0) as caso165, if(total=1650000
AND total<1660000,1,0) as caso166, if(total=1660000 AND total<1670000,1,0) as caso167, if(total=1670000
AND total<1680000,1,0) as caso168, if(total=1680000 AND total<1690000,1,0) as caso169, if(total=1690000
AND total<1700000,1,0) as caso170, if(total=1700000 AND total<1710000,1,0) as caso171, if(total=1710000
AND total<1720000,1,0) as caso172, if(total=1720000 AND total<1730000,1,0) as caso173, if(total=1730000
AND total<1740000,1,0) as caso174, if(total=1740000 AND total<1750000,1,0) as caso175, if(total=1750000
AND total<1760000,1,0) as caso176, if(total=1760000 AND total<1770000,1,0) as caso177, if(total=1770000
AND total<1780000,1,0) as caso178, if(total=1780000 AND total<1790000,1,0) as caso179, if(total=1790000
AND total<1800000,1,0) as caso180, if(total=1800000 AND total<1810000,1,0) as caso181, if(total=1810000
AND total<1820000,1,0) as caso182, if(total=1820000 AND total<1830000,1,0) as caso183, if(total=1830000
AND total<1840000,1,0) as caso184, if(total=1840000 AND total<1850000,1,0) as caso185, if(total=1850000
AND total<1860000,1,0) as caso186, if(total=1860000 AND total<1870000,1,0) as caso187, if(total=1870000
AND total<1880000,1,0) as caso188, if(total=1880000 AND total<1890000,1,0) as caso189, if(total=1890000
AND total<1900000,1,0) as caso190, if(total=1900000 AND total<1910000,1,0) as caso191, if(total=1910000
AND total<1920000,1,0) as caso192, if(total=1920000 AND total<1930000,1,0) as caso193, if(total=1930000
AND total<1940000,1,0) as caso194, if(total=1940000 AND total<1950000,1,0) as caso195, if(total=1950000
AND total<1960000,1,0) as caso196, if(total=1960000 AND total<1970000,1,0) as caso197, if(total=1970000
AND total<1980000,1,0) as caso198, if(total=1980000 AND total<1990000,1,0) as caso199, if(total=1990000
AND total<2000000,1,0) as caso200, if(total=2000000 AND total<2010000,1,0) as caso201, if(total=2010000
AND total<2020000,1,0) as caso202, if(total=2020000 AND total<2030000,1,0) as caso203, if(total=2030000
AND total<2040000,1,0) as caso204, if(total=2040000 AND total<2050000,1,0) as caso205, if(total=2050000
AND total<2060000,1,0) as caso206, if(total=2060000 AND total<2070000,1,0) as caso207, if(total=2070000
AND total<2080000,1,0) as caso208, if(total=2080000 AND total<2090000,1,0) as caso209, if(total=2090000
AND total<2100000,1,0) as caso210, if(total=2100000 AND total<2110000,1,0) as caso211, if(total=2110000
AND total<2120000,1,0) as caso212, if(total=2120000 AND total<2130000,1,0) as caso213, if(total=2130000
AND total<2140000,1,0) as caso214, if(total=2140000 AND total<2150000,1,0) as caso215, if(total=2150000
AND total<2160000,1,0) as caso216, if(total=2160000 AND total<2170000,1,0) as caso217, if(total=2170000
AND total<2180000,1,0) as caso218, if(total=2180000 AND total<2190000,1,0) as caso219, if(total=2190000
AND total<2200000,1,0) as caso220, if(total=2200000 AND total<2210000,1,0) as caso221, if(total=2210000
AND total<2220000,1,0) as caso222, if(total=2220000 AND total<2230000,1,0) as caso223, if(total=2230000
AND total<2240000,1,0) as caso224, if(total=2240000 AND total<2250000,1,0) as caso225, if(total=2250000
AND total<2260000,1,0) as caso226, if(total=2260000 AND total<2270000,1,0) as caso227, if(total=2270000
AND total<2280000,1,0) as caso228, if(total=2280000 AND total<2290000,1,0) as caso229, if(total=2290000
AND total<2300000,1,0) as caso230, if(total=2300000 AND total<2310000,1,0) as caso231, if(total=2310000
AND total<2320000,1,0) as caso232, if(total=2320000 AND total<2330000,1,0) as caso233, if(total=2330000
AND total<2340000,1,0) as caso234, if(total=2340000 AND total<2350000,1,0) as caso235, if(total=2350000
AND total<2360000,1,0) as caso236, if(total=2360000 AND total<2370000,1,0) as caso237, if(total=2370000
AND total<2380000,1,0) as caso238, if(total=2380000 AND total<2390000,1,0) as caso239, if(total=2390000
AND total<2400000,1,0) as caso240, if(total=2400000 AND total<2410000,1,0) as caso241, if(total=2410000
AND total<2420000,1,0) as caso242, if(total=2420000 AND total<2430000,1,0) as caso243, if(total=2430000
AND total<2440000,1,0) as caso244, if(total=2440000 AND total<2450000,1,0) as caso245, if(total=2450000
AND total<2460000,1,0) as caso246, if(total=2460000 AND total<2470000,1,0) as caso247, if(total=2470000
AND total<2480000,1,0) as caso248, if(total=2480000 AND total<2490000,1,0) as caso249, if(total=2490000
AND total<2500000,1,0) as caso250, if(total=2500000 AND total<2510000,1,0) as caso251, if(total=2510000
AND total<2520000,1,0) as caso252, if(total=2520000 AND total<2530000,1,0) as caso253, if(total=2530000
AND total<2540000,1,0) as caso254, if(total=2540000 AND total<2550000,1,0) as caso255, if(total=2550000
AND total<2560000,1,0) as caso256, if(total=2560000 AND total<2570000,1,0) as caso257, if(total=2570000
AND total<2580000,1,0) as caso258, if(total=2580000 AND total<2590000,1,0) as caso259, if(total=2590000
AND total<2600000,1,0) as caso260, if(total=2600000 AND total<2610000,1,0) as caso261, if(total=2610000
AND total<2620000,1,0) as caso262, if(total=2620000 AND total<2630000,1,0) as caso263, if(total=2630000
AND total<2640000,1,0) as caso264, if(total=2640000 AND total<2650000,1,0) as caso265, if(total=2650000
AND total<2660000,1,0) as caso266, if(total=2660000 AND total<2670000,1,0) as caso267, if(total=2670000
AND total<2680000,1,0) as caso268, if(total=2680000 AND total<2690000,1,0) as caso269, if(total=2690000
AND total<2700000,1,0) as caso270, if(total=2700000 AND total<2710000,1,0) as caso271, if(total=2710000
AND total<2720000,1,0) as caso272, if(total=2720000 AND total<2730000,1,0) as caso273, if(total=2730000
AND total<2740000,1,0) as caso274, if(total=2740000 AND total<2750000,1,0) as caso275, if(total=2750000
AND total<2760000,1,0) as caso276, if(total=2760000 AND total<2770000,1,0) as caso277, if(total=2770000
AND total<2780000,1,0) as caso278, if(total=2780000 AND total<2790000,1,0) as caso279, if(total=2790000
AND total<2800000,1,0) as caso280, if(total=2800000 AND total<2810000,1,0) as caso281, if(total=2810000
AND total<2820000,1,0) as caso282, if(total=2820000 AND total<2830000,1,0) as caso283, if(total=2830000
AND total<2840000,1,0) as caso284, if(total=2840000 AND total<2850000,1,0) as caso285, if(total=2850000
AND total<2860000,1,0) as caso286, if(total=2860000 AND total<2870000,1,0) as caso287, if(total=2870000
AND total<2880000,1,0) as caso288, if(total=2880000 AND total<2890000,1,0) as caso289, if(total=2890000
AND total<2900000,1,0) as caso290, if(total=2900000 AND total<2910000,1,0) as caso291, if(total=2910000
AND total<2920000,1,0) as caso292, if(total=2920000 AND total<2930000,1,0) as caso293, if(total=2930000
AND total<2940000,1,0) as caso294, if(total=2940000 AND total<2950000,1,0) as caso295, if(total=2950000
AND total<2960000,1,0) as caso296, if(total=2960000 AND total<2970000,1,0) as caso297, if(total=2970000
AND total<2980000,1,0) as caso298, if(total=2980000 AND total<2990000,1,0) as caso299, if(total=2990000
AND total<3000000,1,0) as caso300, if(total=3000000 AND total<3010000,1,0) as caso301, if(total=3010000
AND total<3020000,1,0) as caso302, if(total=3020000 AND total<3030000,1,0) as caso303, if(total=3030000
AND total<3040000,1,0) as caso304, if(total=3040000 AND total<3050000,1,0) as caso305, if(total=3050000
AND total<3060000,1,0) as caso306, if(total=3060000 AND total<3070000,1,0) as caso307, if(total=3070000
AND total<3080000,1,0) as caso308, if(total=3080000 AND total<3090000,1,0) as caso309, if(total=3090000
AND total<3100000,1,0) as caso310, if(total=3100000 AND total<3110000,1,0) as caso311, if(total=3110000
AND total<3120000,1,0) as caso312, if(total=3120000 AND total<3130000,1,0) as caso313, if(total=3130000
AND total<3140000,1,0) as caso314, if(total=3140000 AND total<3150000,1,0) as caso315, if(total=3150000
AND total<3160000,1,0) as caso316, if(total=3160000 AND total<3170000,1,0) as caso317, if(total=3170000
AND total<3180000,1,0) as caso318, if(total=3180000 AND total<3190000,1,0) as caso319, if(total=3190000
AND total<3200000,1,0) as caso320, if(total=3200000 AND total<3210000,1,0) as caso321, if(total=3210000
AND total<3220000,1,0) as caso322, if(total=3220000 AND total<3230000,1,0) as caso323, if(total=3230000
AND total<3240000,1,0) as caso324, if(total=3240000 AND total<3250000,1,0) as caso325, if(total=3250000
AND total<3260000,1,0) as caso326, if(total=3260000 AND total<3270000,1,0) as caso327, if(total=3270000
AND total<3280000,1,0) as caso328, if(total=3280000 AND total<3290000,1,0) as caso329, if(total=3290000
AND total<3300000,1,0) as caso330, if(total=3300000 AND total<3310000,1,0) as caso331, if(total=3310000
AND total<3320000,1,0) as caso332, if(total=3320000 AND total<3330000,1,0) as caso333, if(total=3330000
AND total<3340000,1,0) as caso334, if(total=3340000 AND total<3350000,1,0) as caso335, if(total=3350000
AND total<3360000,1,0) as caso336, if(total=3360000 AND total<3370000,1,0) as caso337, if(total=3370000
AND total<3380000,1,0) as caso338, if(total=3380000 AND total<3390000,1,0) as caso339, if(total=3390000
AND total<3400000,1,0) as caso340, if(total=3400000 AND total<3410000,1,0) as caso341, if(total=3410000
AND total<3420000,1,0) as caso342, if(total=3420000 AND total<3430000,1,0) as caso343, if(total=3430000
AND total<3440000,1,0) as caso344, if(total=3440000 AND total<3450000,1,0) as caso345, if(total=3450000
AND total<3460000,1,0) as caso346, if(total=3460000 AND total<3470000,1,0) as caso347, if(total=3470000
AND total<3480000,1,0) as caso348, if(total=3480000 AND total<3490000,1,0) as caso349, if(total=3490000
AND total<3500000,1,0) as caso350, if(total=3500000 AND total<3510000,1,0) as caso351, if(total=3510000
AND total<3520000,1,0) as caso352, if(total=3520000 AND total<3530000,1,0) as caso353, if(total=3530000
AND total<3540000,1,0) as caso354, if(total=3540000 AND total<3550000,1,0) as caso355, if(total=3550000
AND total<3560000,1,0) as caso356, if(total=3560000 AND total<3570000,1,0) as caso357, if(total=3570000
AND total<3580000,1,0) as caso358, if(total=3580000 AND total<3590000,1,0) as caso359, if(total=3590000
AND total<3600000,1,0) as caso360, if(total=3600000 AND total<3610000,1,0) as caso361, if(total=3610000
AND total<3620000,1,0) as caso362, if(total=3620000 AND total
```

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:

```
mysql> source /media/sf_Compartida/script1.sql
Database changed
Query OK, 0 rows affected, 1 warning (0.01 sec)
Query OK, 0 rows affected, 1 warning (0.01 sec)
Query OK, 0 rows affected, 1 warning (0.01 sec)
Query OK, 0 rows affected, 1 warning (0.00 sec)
Query OK, 0 rows affected (0.00 sec)
Query OK, 0 rows affected (0.00 sec)

+----+
| Type | Metric | Metric | Metric |
+----+
| Sales | 1000 | 1000000 | 1000000 |
+----+
1 row in set (0.00 sec)

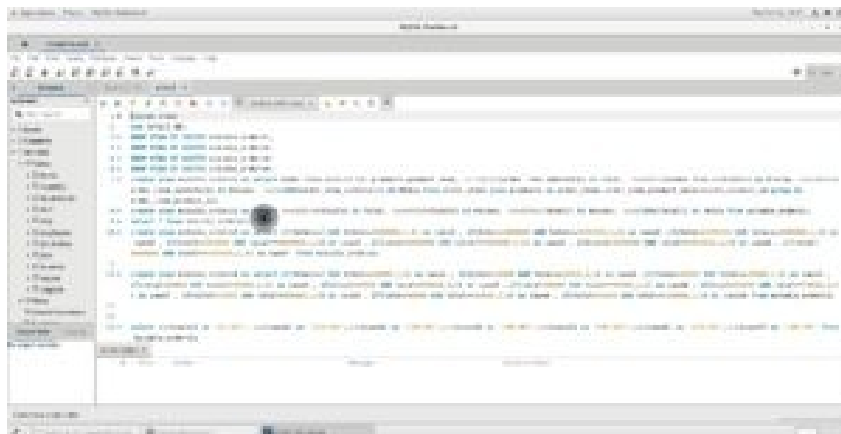
Query OK, 0 rows affected (0.00 sec)
Query OK, 0 rows affected (0.00 sec)

+----+
| ID | Sales | Sales | Sales | Sales | Sales | Sales | Sales | Sales | Sales |
+----+
| 1 | 100 | 1000 | 10000 | 100000 | 1000000 | 10000000 | 100000000 | 1000000000 | 10000000000 |
+----+
1 row in set (0.00 sec)
```

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](#).

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](#). Y las tools par VMware, se pueden descargar [aquí](#).

En [este post](#) 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:







CDH_6.3.2_CentOS7

 Power on this virtual machine

 Edit virtual machine settings

 Upgrade this virtual machine

▼ Devices

 Memory	12 GB
 Processors	4
 Hard Disk (SCSI)	500 GB
 Network Adapter	Custom (VMnet8)
 USB Controller	Present
 Display	Auto detect

▼ Description

Type here to enter a description of this virtual machine.

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

```
osboxes@quickstart-bigdata:~$ ifconfig
bash: ifconfig: command not found...
osboxes@quickstart-bigdata:~$ ifconfig
ens33: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
    inet 10.10.1.3 netmask 255.255.255.0 broadcast 10.10.1.255
    inet6 fe80::be92:b62f:fbfc:d974 prefixlen 64 scopeid 0x20<link>
    ether 08:0c:29:8b:45:f2 txqueuelen 1000 (Ethernet)
    RX packets 1499 bytes 398407 (389.1 KiB)
    RX errors 0 dropped 0 overruns 0 frame 0
    TX packets 326 bytes 35142 (34.3 KiB)
    TX errors 0 dropped 0 overruns 0 carrier 0 collisions 0
```

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.

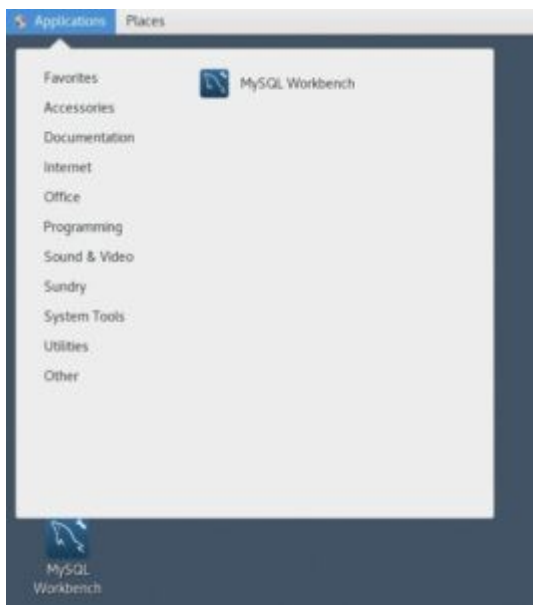
```
sudo yum install epel-release -y
```

```
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:



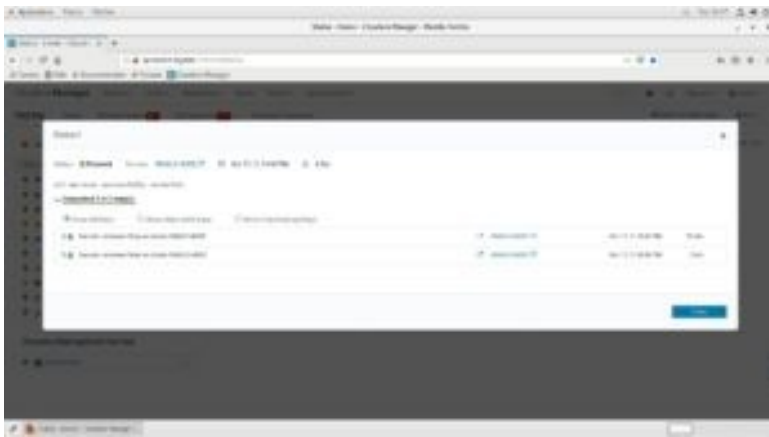
Pero a veces puede ocurrir que muchos servicios se presenten en rojo, tal como la siguiente imagen. En ese caso, en el desplegable que hay al lado de “single-node”, seleccionamos “restart”:



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.

