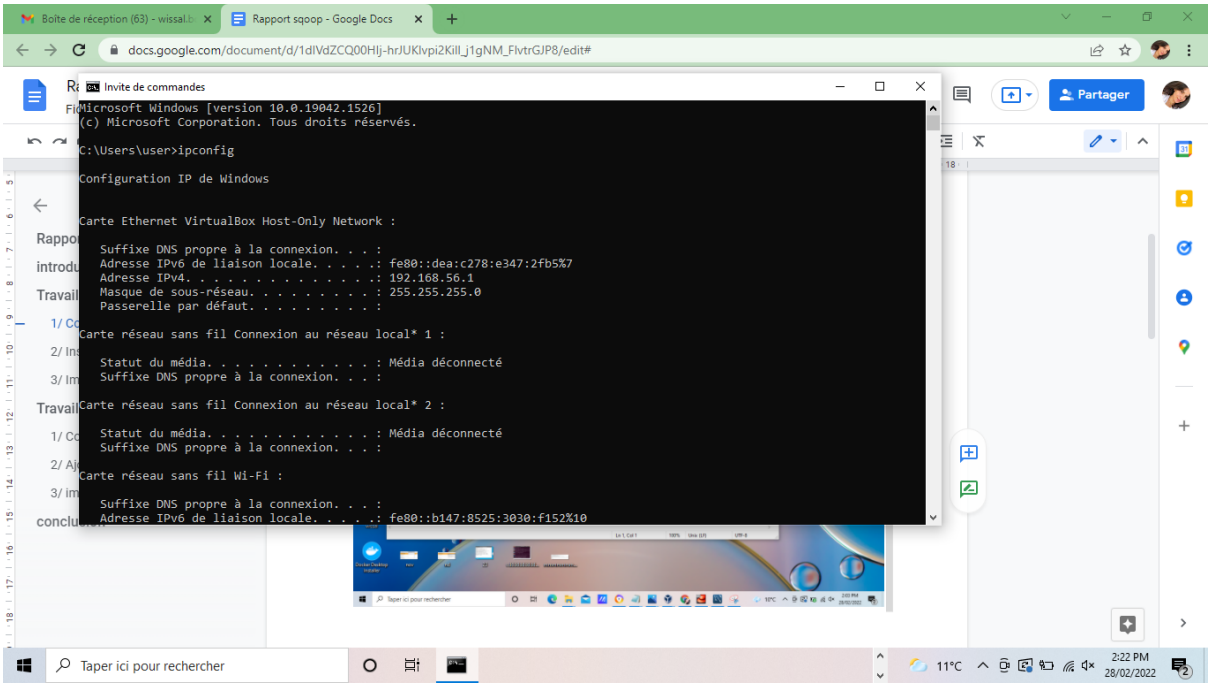
Rapport Sqoop

Introduction

Dans ce mini-projet, nous allons essayer d'importer une base de données mysql dans notre machine hôte vers la machine virtuelle Cloudera, en utilisant l'outil Sqoop.

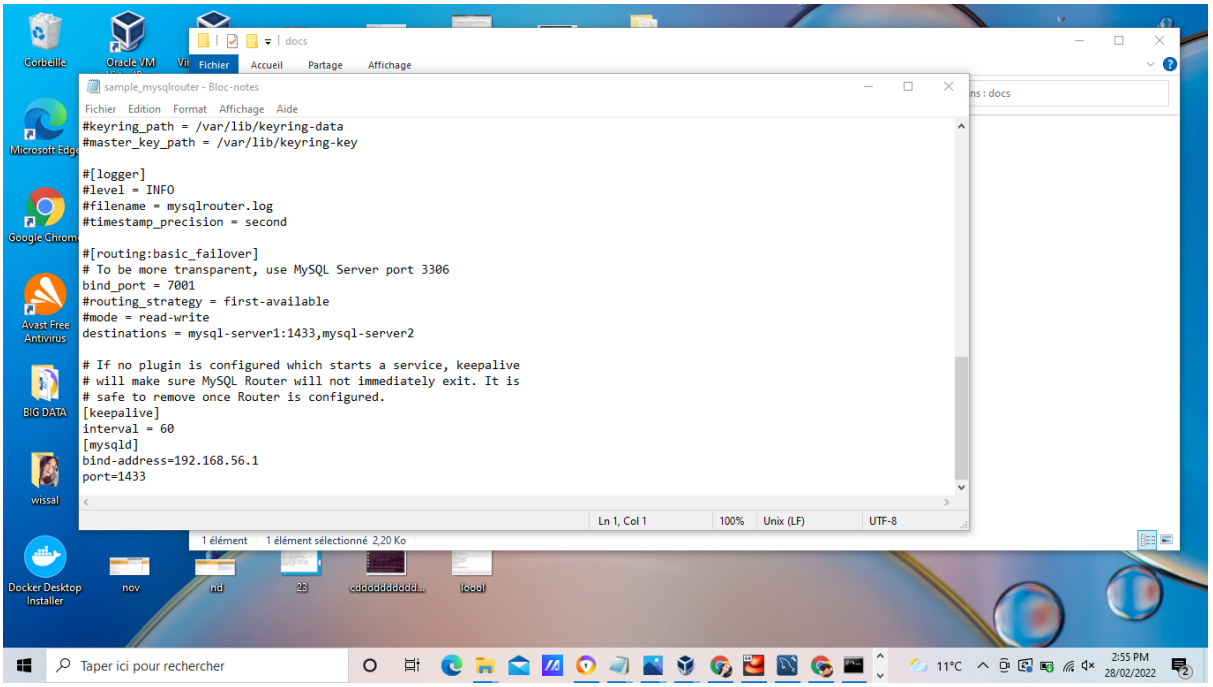
Travail réalisé sur la machine hôte Windows:

1/ Détermination de l'adresse ip du host

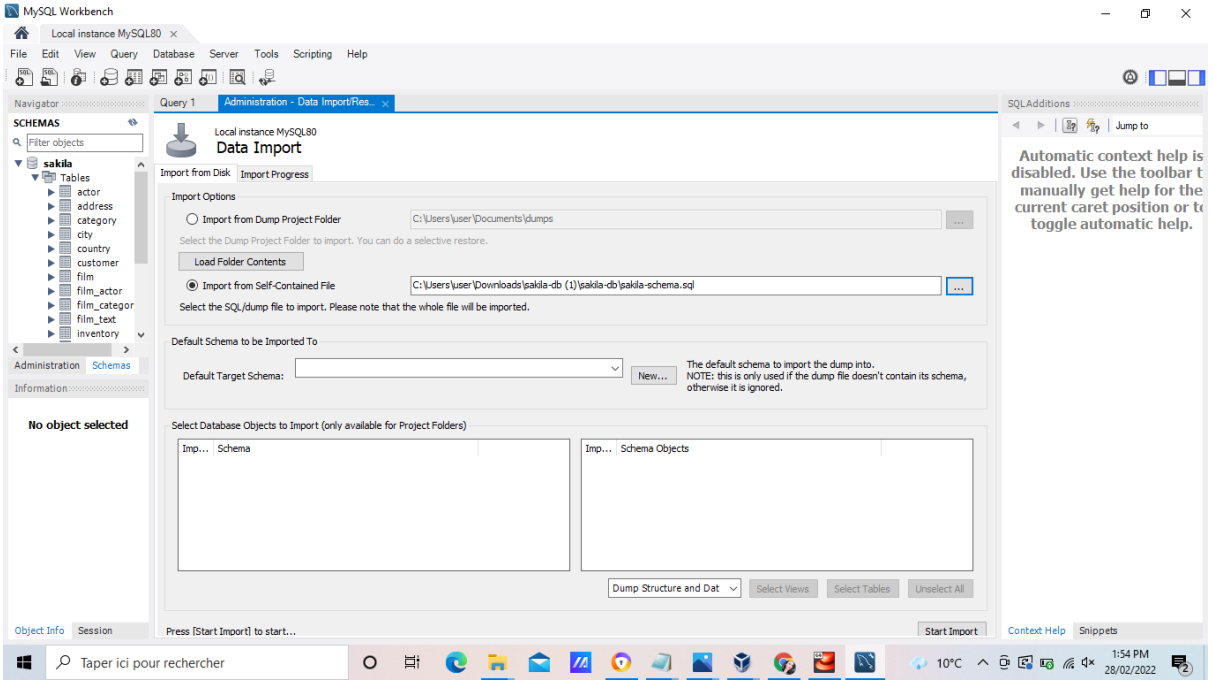


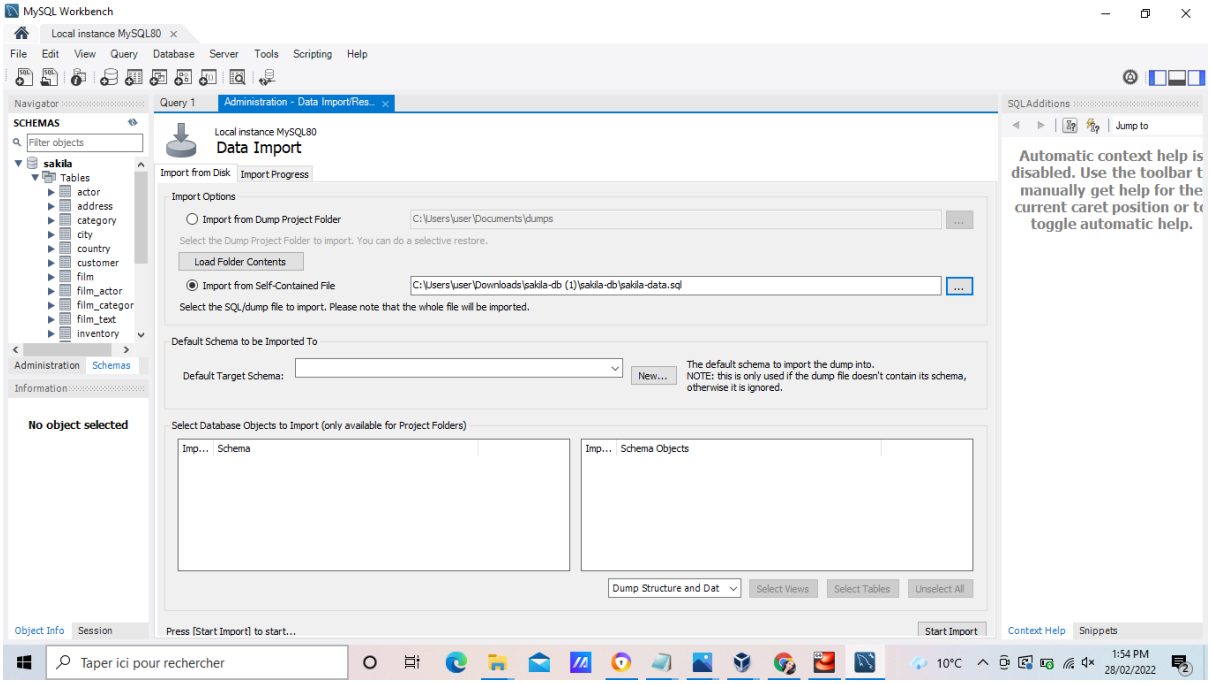
2/ Modification du fichier C:\Program Files/MYSQL/MYSQLRouter8.0/docs/sample_mysqlrouter

Ajout de la section mysqld dans laquelle on définit:
-l'élément bind-address: 192.168.56.1
-l'élément port :1433

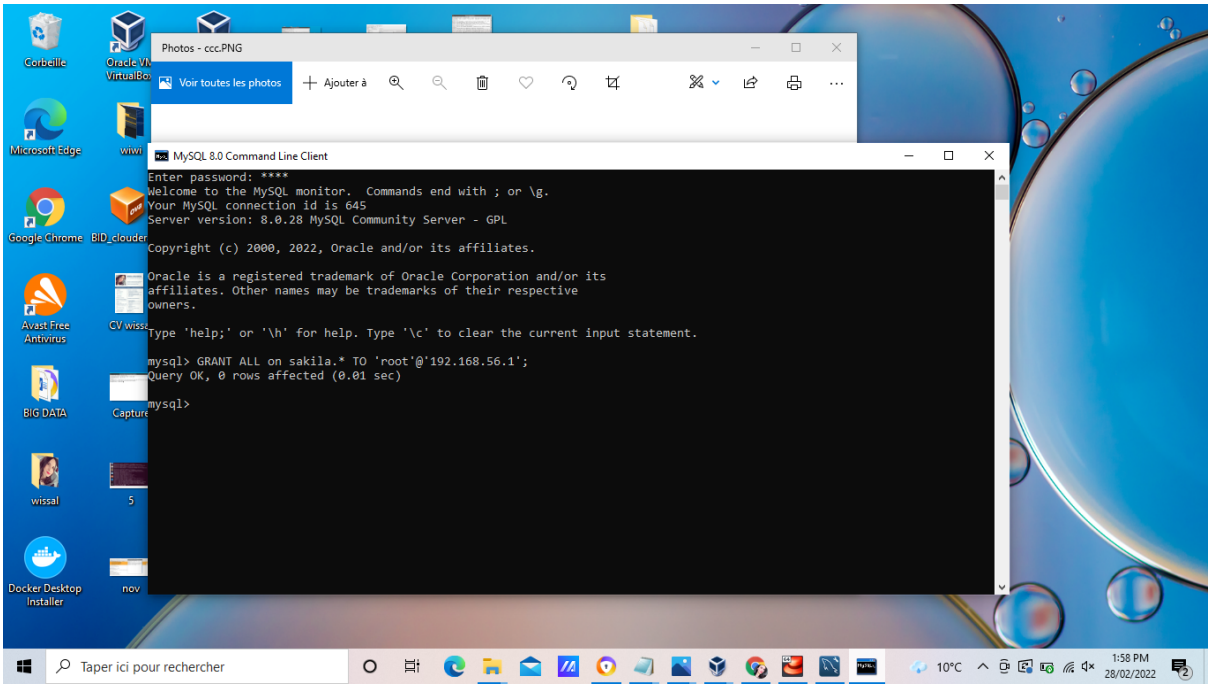


3/ Import de la base de données mysql Sakila sur mySQLWorkbench

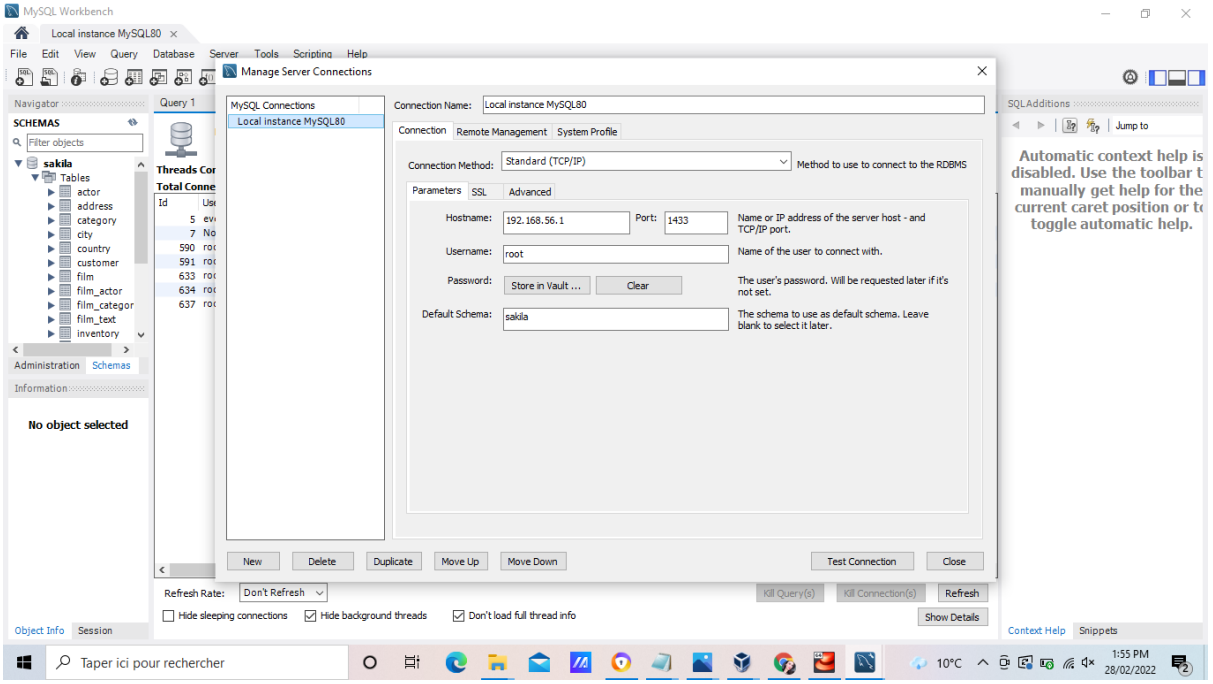



MySQL Workbench
Local instance MySQL80
File Edit View Query Database Server Tools Scripting Help
Navigator
SCHEMAS
Filter objects
sakila
Tables
actor
address
category
city
country
customer
film
film_actor
film_categor
film_text
inventory
Administration Schemas
Information
No object selected
Object Info Session
Query 1
Administration - Data Import/Res...
Local instance MySQL80
Data Import
Import from Disk
Import Progress
Import Options
Import from Dump Project Folder
C:\Users\user\Documents\dumps
Select the Dump Project Folder to import. You can do a selective restore.
Load Folder Contents
Import from Self-Contained File
C:\Users\user\Downloads\sakila-db (1)\sakila-db\sakila-data.sql
Select the SQL/dump file to import. Please note that the whole file will be imported.
Default Schema to be Imported To
Default Target Schema:
New...
The default schema to import the dump into.
NOTE: this is only used if the dump file doesn't contain its schema, otherwise it is ignored.
Select Database Objects to Import (only available for Project Folders)
Imp... Schema
Imp... Schema Objects
Dump Structure and Dat
Select Views
Select Tables
Unselect All
Press [Start Import] to start...
Start Import
SQLAdditions
Jump to
Automatic context help is disabled. Use the toolbar t manually get help for the current caret position or t toggle automatic help.
Context Help Snippets
Taper ici pour rechercher
10°C
1:54 PM
28/02/2022

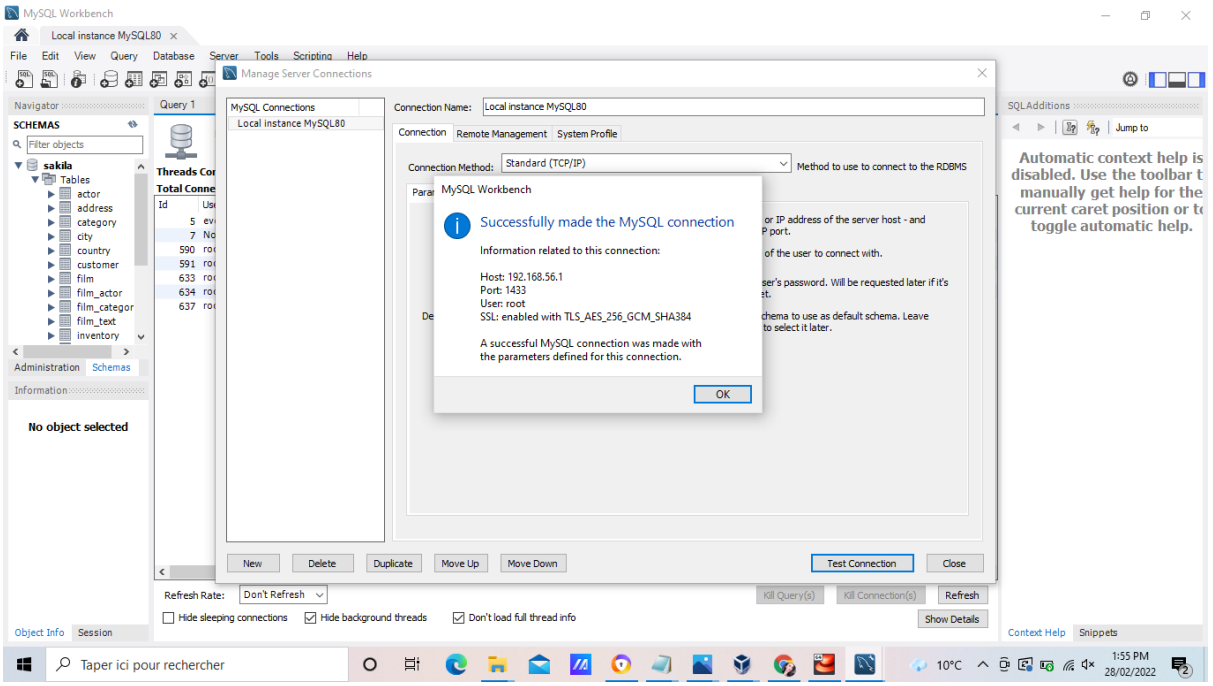
4/ Donner les privilèges à l'utilisateur root sur la base Sakila :



5/ Configuration de la connexion au serveur



6/ Configuration de l’instance locale

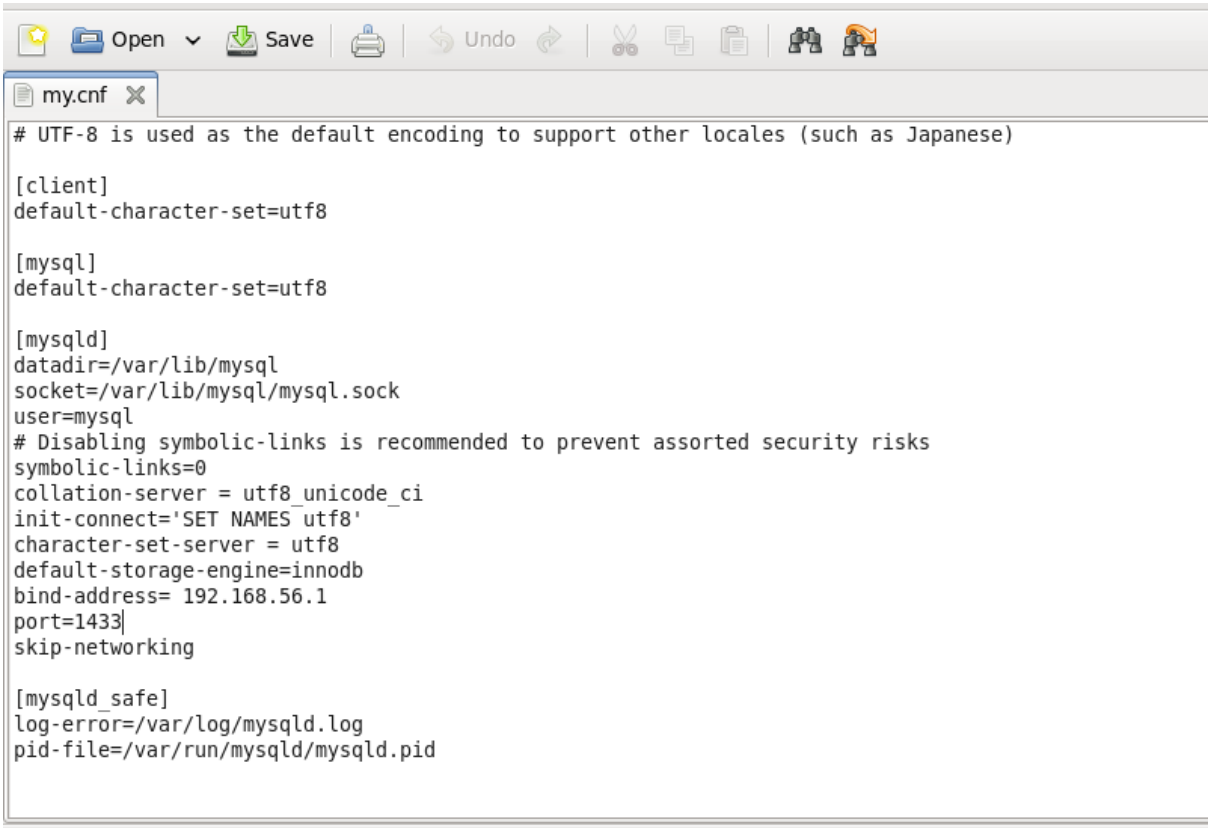


Travail réalisé sur la machine virtuelle Cloudera

1/ Configuration du fichier my.cnf :

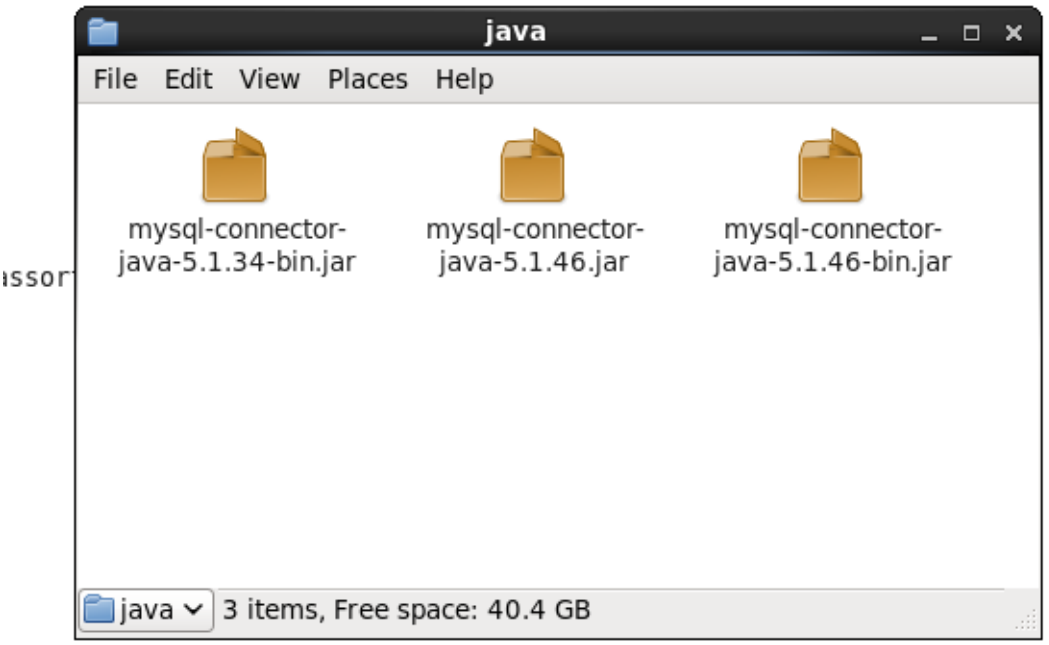
Dans la section [mysqld]:

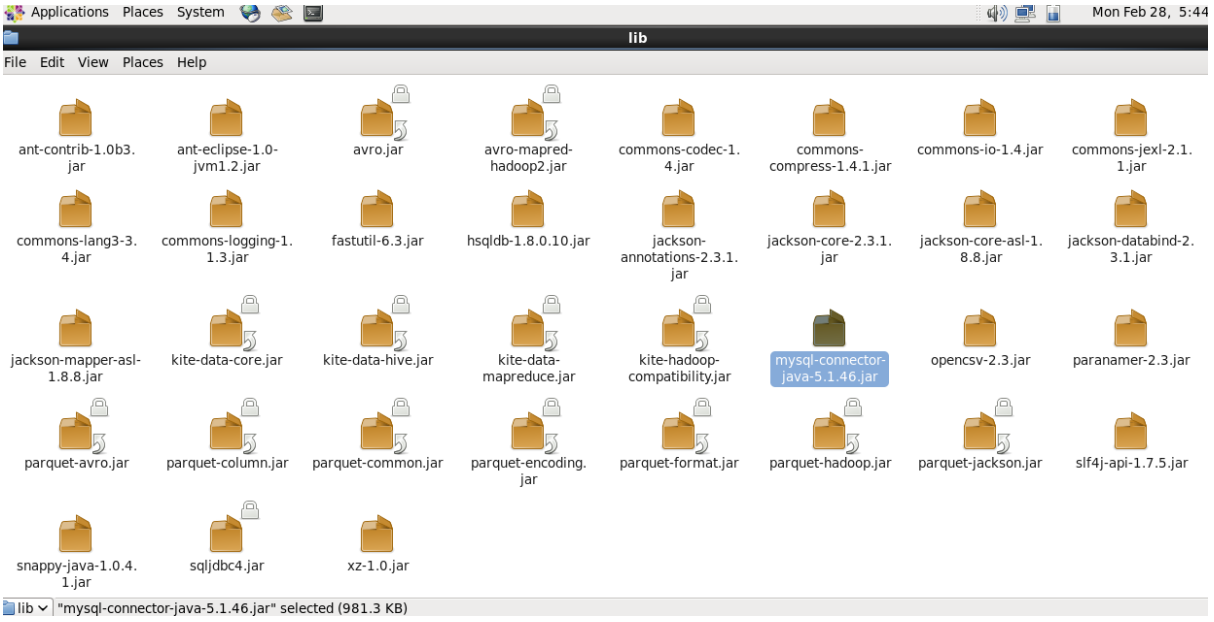
-Ajout de bind address: 192.168.56.1
-Ajout du port: 1433



2/ Téléchargement et ajout de mysqlconnector.jar dans :

Il faut ajouter les deux fichiers .jar du connecteur mysql pour pouvoir faire la connexion depuis Cloudera à MYSQLServer dans windows dans
l'emplacement: usr/share/JAVA
l'emplacement: /usr/lib/sqoop/lib


java
File Edit View Places Help
mysql-connector-java-5.1.34-bin.jar
mysql-connector-java-5.1.46.jar
mysql-connector-java-5.1.46-bin.jar
java
3 items, Free space: 40.4 GB

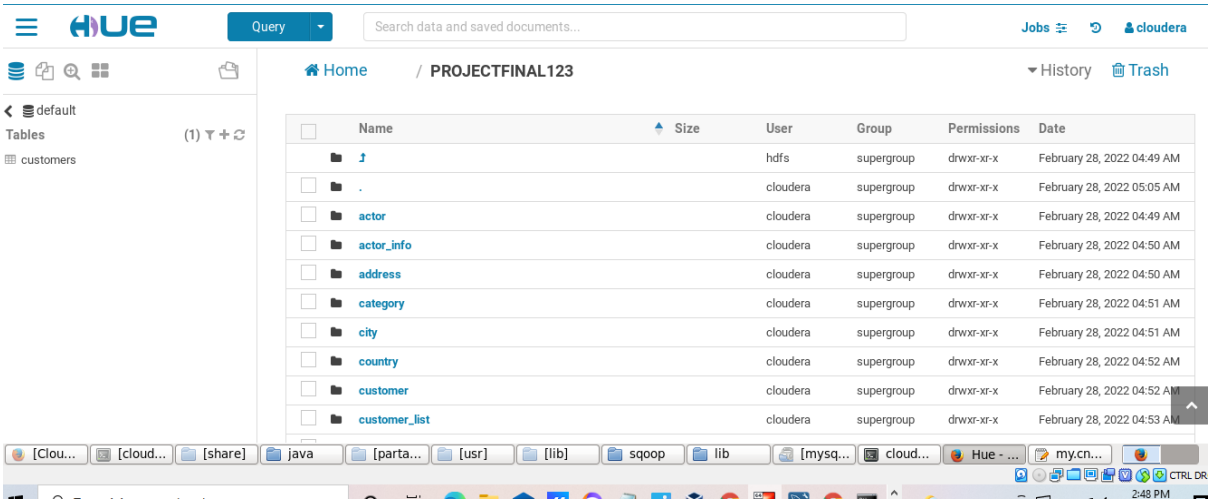

Applications Places System
Mon Feb 28, 5:44
lib
File Edit View Places Help
ant-contrib-1.0b3.jar
ant-eclipse-1.0-jvm1.2.jar
avro.jar
avro-mapred-hadoop2.jar
commons-codec-1.4.jar
commons-compress-1.4.1.jar
commons-io-1.4.jar
commons-jexl-2.1.1.jar
commons-lang3-3.4.jar
commons-logging-1.1.3.jar
fastutil-6.3.jar
hsqldb-1.8.0.10.jar
jackson-annotations-2.3.1.jar
jackson-core-2.3.1.jar
jackson-core-asl-1.8.8.jar
jackson-databind-2.3.1.jar
jackson-mapper-asl-1.8.8.jar
kite-data-core.jar
kite-data-hive.jar
kite-data-mapreduce.jar
kite-hadoop-compatibility.jar
mysql-connector-java-5.1.46.jar
opencsv-2.3.jar
paranamer-2.3.jar
parquet-avro.jar
parquet-column.jar
parquet-common.jar
parquet-encoding.jar
parquet-format.jar
parquet-hadoop.jar
parquet-jackson.jar
slf4j-api-1.7.5.jar
snappy-java-1.0.4.1.jar
sqljdbc4.jar
xz-1.0.jar
lib
"mysql-connector-java-5.1.46.jar" selected (981.3 KB)

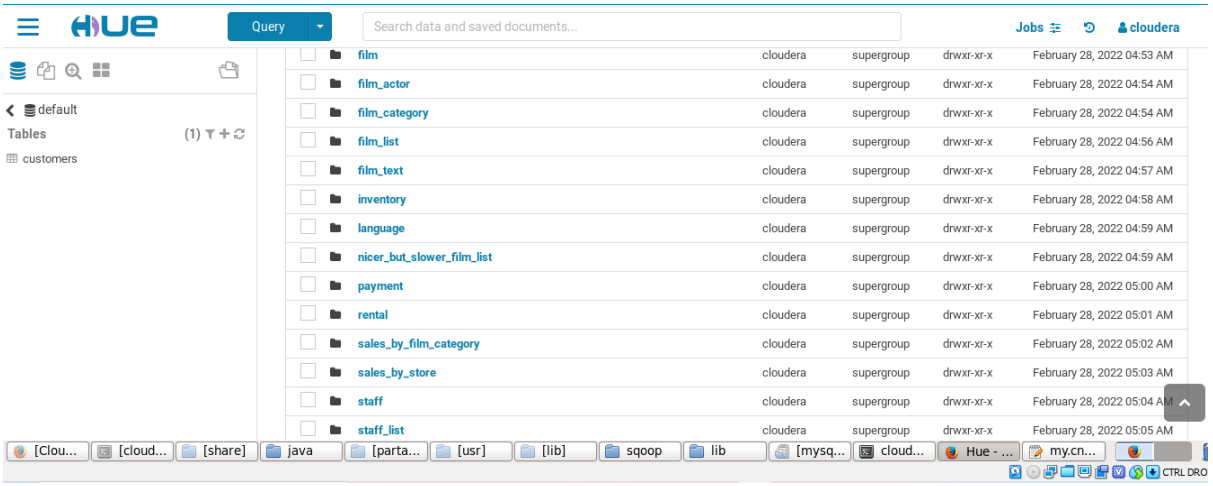
3/ import avec sqoop

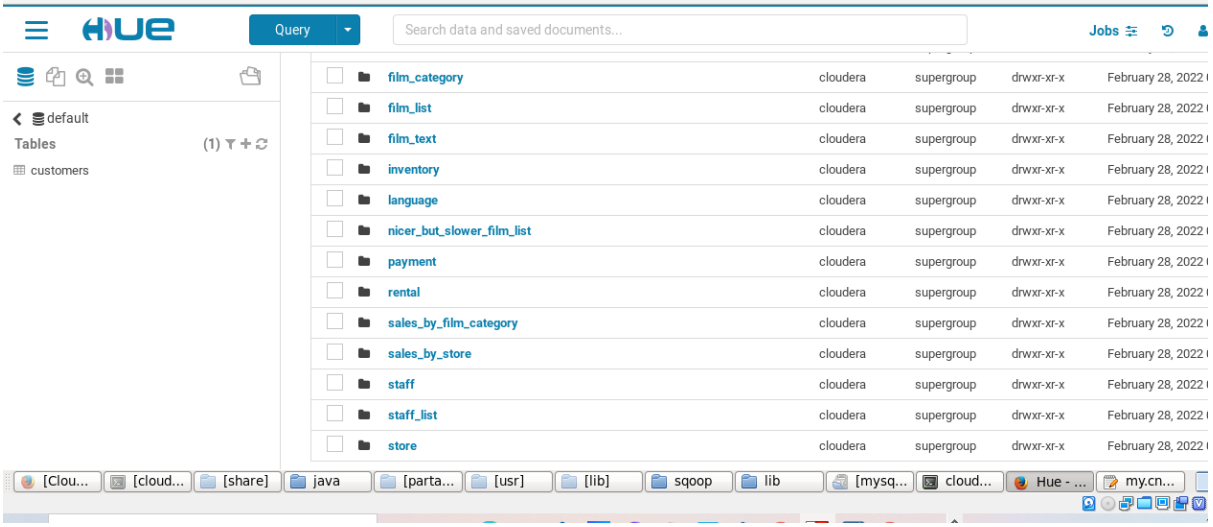
```
                Virtual memory (bytes) snapshot=0
22/02/28 04:31:51 WARN mapreduce.Counters: Group FileSystemCounters is deprecated. Use org.apache.hadoop.mapreduce.FileSystemCounter instead
22/02/28 04:31:51 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 133.5952 seconds (0 bytes/sec)
22/02/28 04:31:51 INFO mapreduce.ImportJobBase: Retrieved 0 records.
22/02/28 04:31:51 ERROR tool.ImportAllTablesTool: Error during import: Import job failed!
[cloudera@quickstart ~]$ sqoop list-databases --connect jdbc:mysql://192.168.56.1:1433  --username root --password root
Warning: /usr/lib/sqoop/../accumulo does not exist! Accumulo imports will fail.
Please set $ACCUMULO_HOME to the root of your Accumulo installation.
22/02/28 04:46:26 INFO sqoop.Sqoop: Running Sqoop version: 1.4.6-cdh5.12.0
22/02/28 04:46:26 WARN tool.BaseSqoopTool: Setting your password on the command-line is insecure. Consider using -P instead.
22/02/28 04:46:27 INFO manager.MySQLManager: Preparing to use a MySQL streaming resultset.
Mon Feb 28 04:46:28 PST 2022 WARN: Establishing SSL connection without server's identity verification is not recommended. According to MySQL 5.5.45+, 5.6.26+ and 5.7.6+
 requirements SSL connection must be established by default if explicit option isn't set. For compliance with existing applications not using SSL the verifyServerCertif
icate property is set to 'false'. You need either to explicitly disable SSL by setting useSSL=false, or set useSSL=true and provide truststore for server certificate ve
rification.
mysql
information_schema
performance_schema
sys
sakila
[cloudera@quickstart ~]$
```

```
cloudera@quickstart:~
File  Edit  View  Search  Terminal  Help
                HDFS: Number of bytes read=87
                HDFS: Number of bytes written=56
                HDFS: Number of read operations=4
                HDFS: Number of large read operations=0
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=1
                Other local map tasks=1
                Total time spent by all maps in occupied slots (ms)=21151
                Total time spent by all reduces in occupied slots (ms)=0
                Total time spent by all map tasks (ms)=21151
                Total vcore-milliseconds taken by all map tasks=21151
                Total megabyte-milliseconds taken by all map tasks=21658624
        Map-Reduce Framework
                Map input records=2
                Map output records=2
                Input split bytes=87
                Spilled Records=0
                Failed Shuffles=0
                Merged Map outputs=0
                GC time elapsed (ms)=377
                CPU time spent (ms)=3050
                Physical memory (bytes) snapshot=137240576
                Virtual memory (bytes) snapshot=1510211584
                Total committed heap usage (bytes)=60751872
        File Input Format Counters
                Bytes Read=0
        File Output Format Counters
                Bytes Written=56
2/02/28 05:06:09 INFO mapreduce.ImportJobBase: Transferred 56 bytes in 60.2291 seconds (0.9298 bytes/sec)
2/02/28 05:06:09 INFO mapreduce.ImportJobBase: Retrieved 2 records.
cloudera@quickstart ~]$ sqoop import-all-tables --connect jdbc:mysql://192.168.56.1:1433/sakila  --username root --password root --warehouse-dir /PROJECTFINAL123 -m 1
```

4/ Visualisation du résultat dans Hue







Conclusion:

Ce travail nous a permis d'implémenter un use case réel utile dans la vie professionnelle, et de comprendre comment importer nos données entre différentes machines, qu'elles soient virtuelles ou physiques (avec câblage adéquat).