## 



Big Data

06/12/2024

Jon Fernandez de Gorostiza

[**Dataset 2**](#_au51mny0sx6)

[**Cluster 4**](#_ay1zc3v0brpi)

[**HUE 5**](#_olmhxgth32da)

[**Manual de CRUD 6**](#_5kt7p155srao)

[**HIVE 7**](#_kxntaba03fwi)

[CRUD 9](#_w9c78rveq0yf)

[Create 9](#_s46o0u2l8pgf)

[Read 10](#_oh3gkw3iate9)

[Update 10](#_13x3dirw62ul)

[Delete 11](#_boyvqdz176b6)

[**HBASE 12**](#_nbjps3yud7ry)

[CRUD 14](#_ki89u1ohnov)

[Create 14](#_u4q4ojdby0re)

[Read 15](#_fwz1zrdvsfus)

[Update 15](#_wugvnbul5nsg)

[Delete 16](#_xnzrh5suj8lq)

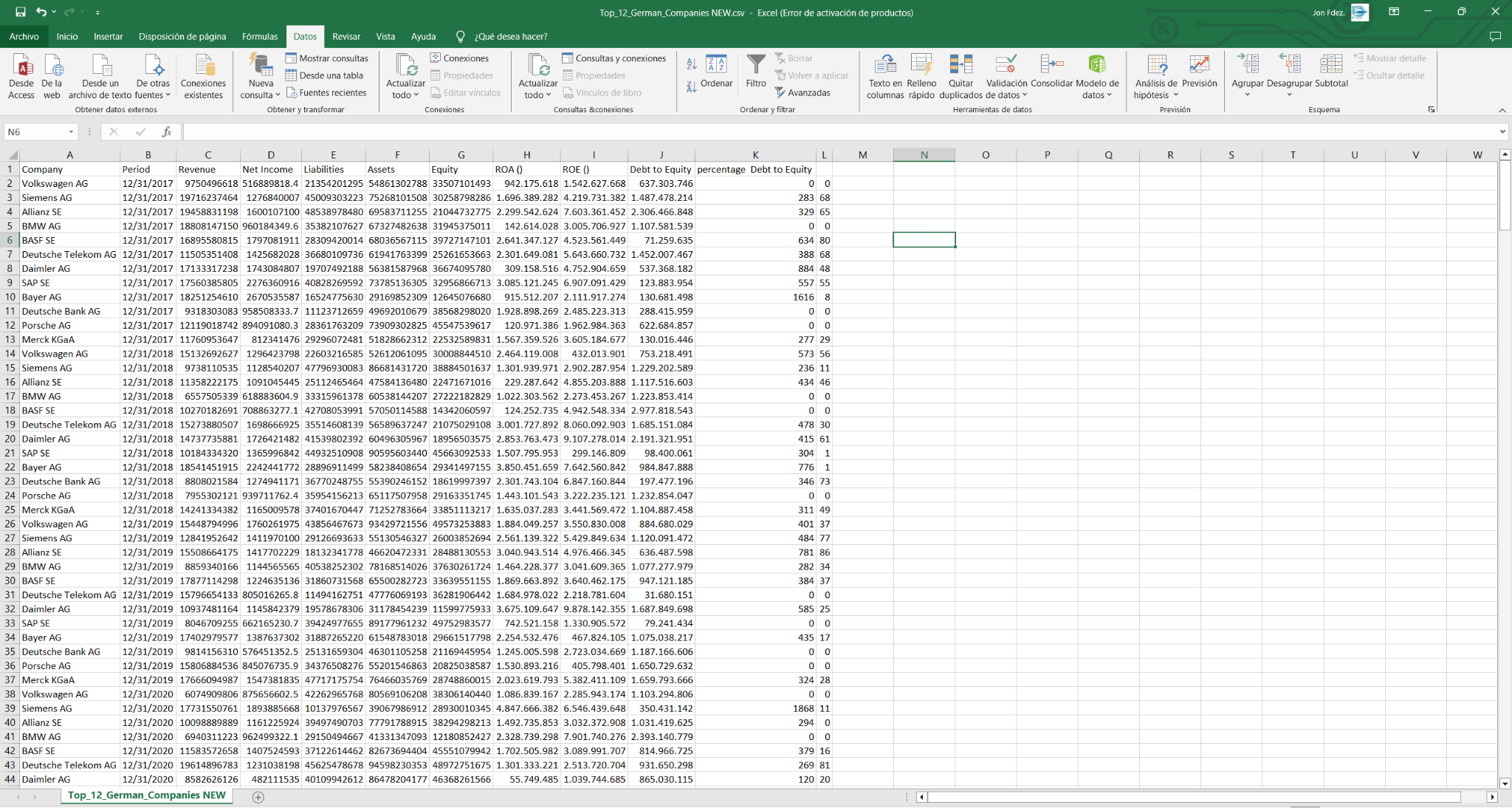
# 

# Dataset

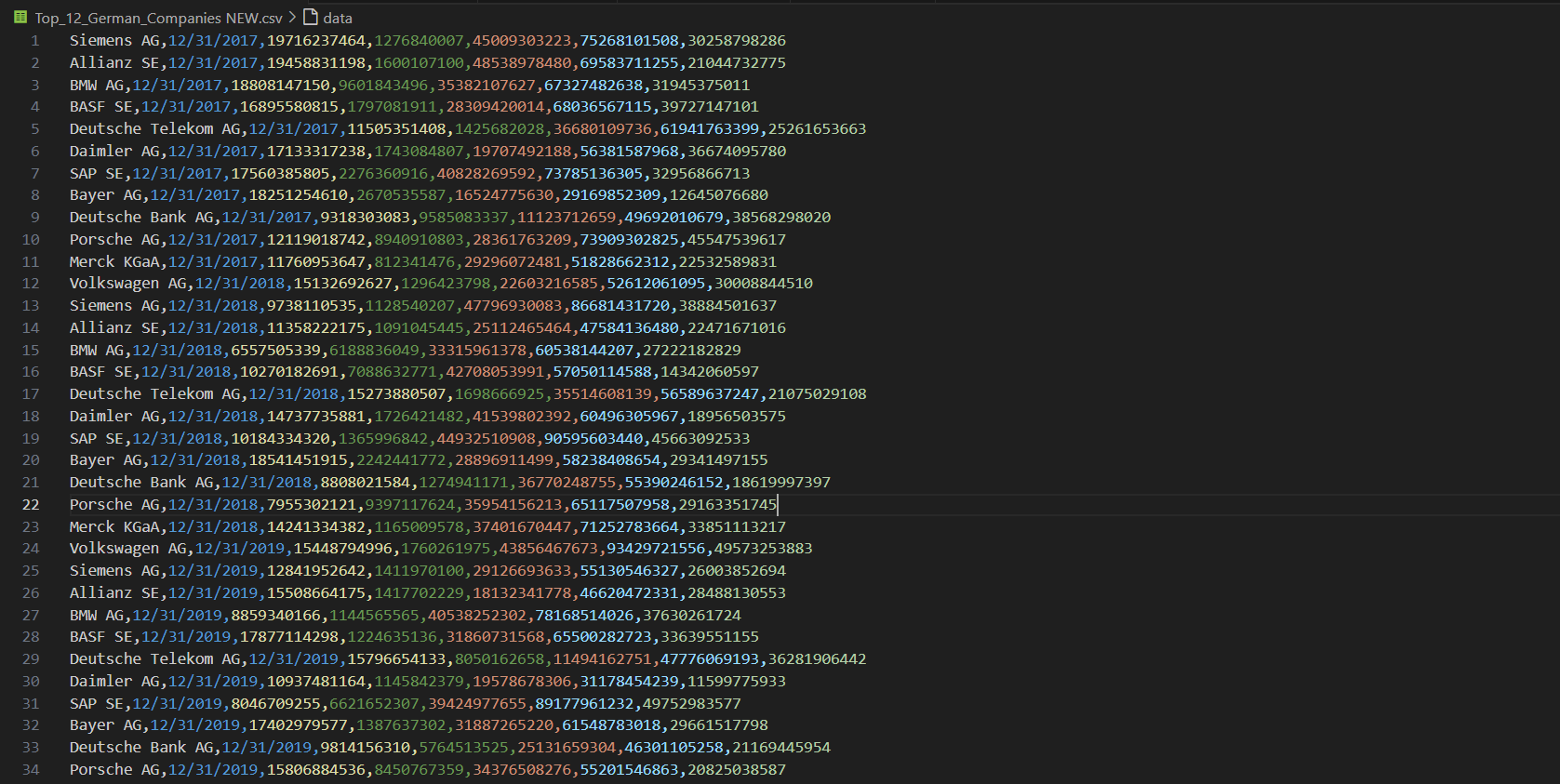
Descargamos el dataset desde Kaggle, en nuestro caso son datos financieros de 12 empresas alemanas.



Al cargar el archivo en excel vemos que hay algunas columnas que como contienen comas, al hacer la división sobre ellas los datos se pasan de columnas. Para solucionar este problema vamos a eliminar la última columna ya que esta es la causante de estos problemas, además de que también contiene el carácter % que nos puede traer mas problemas aun al importar los datos después. Como vamos a crear nosotros la tabla a mano la primera fila, que es la de los nombres de las columnas, la vamos a borrar.



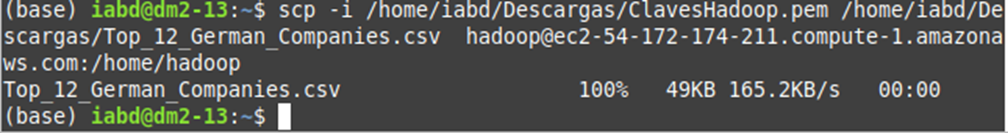
Los datos en formato correcto.



# Cluster

Creamos un clúster de hadoop de la manera que los hemos detallado en los anteriores ejercicios y subimos el archivo csv que hemos descargado de Kaggle al entorno de hadoop.

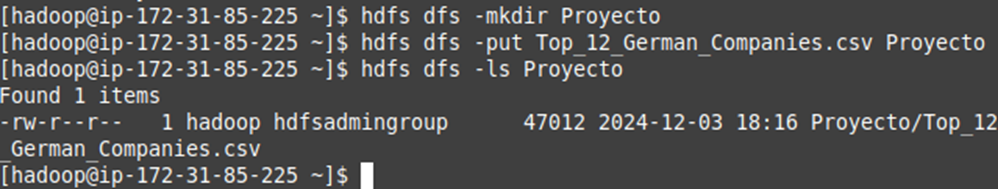
scp -i /home/iabd/Descargas/ClavesHadoop.pem /home/iabd/Descargas/Top\_12\_German\_Companies.csv hadoop@ec2-54-158-105-211.compute-1.amazonaws.com:/home/hadoop



Tenemos que subirlo ahora del entorno al gestor de archivos de hadoop, así que creamos una carpeta y subimos el archivo.

hdfs dfs -mkdir Proyecto

hdfs dfs -put Top\_12\_German\_Companies.csv Proyecto



# HUE

Ahora venimos a aplicaciones del clúster de hadoop en AWS y pinchamos en Tonalidad.

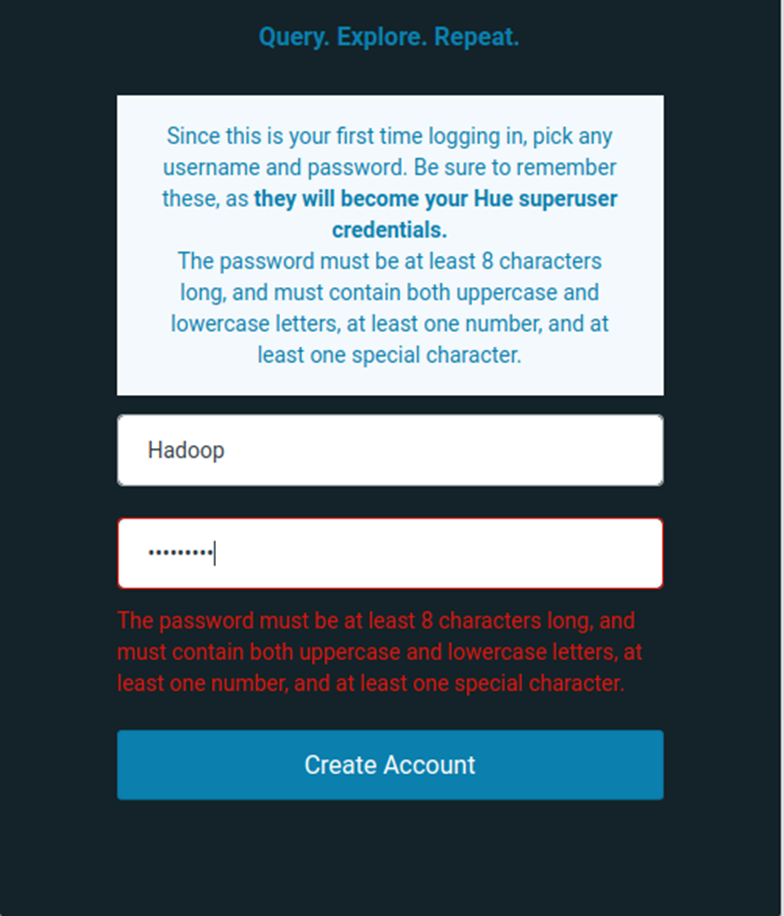


Nos pide crear un usuario.

HUE

User: Hadoop

Password: Hadoop1@



Accedemos a la interfaz de usuario de HUE donde desde esta podemos acceder a HIVE y HBase.

# Manual de CRUD

|  |  |  |
| --- | --- | --- |
| **Operación** | **HBase** | **HIVE** |
| **Create** | create 'table\_name', 'column\_family'  put 'table\_name', 'row\_key', 'column\_family:column\_name', 'value' | CREATE TABLE table\_name (column\_name data\_type, column\_name data\_type)  CREATE TABLE new\_table AS SELECT \* FROM existing\_table |
| **Read** | get 'table\_name', 'row\_key'  scan 'table\_name' | SELECT \* FROM table\_name |
| **Update** | put 'table\_name', 'row\_key', 'column\_family:column\_name', 'new\_value | UPDATE table\_name SET column\_name = new\_value WHERE condition |
| **Delete** | delete 'table\_name', 'row\_key', 'column\_family:column\_name'  deleteall 'table\_name', 'row\_key' | DELETE FROM table\_name WHERE condition |

# HIVE

Desde la interfaz de usuario de HUE, seleccionamos la BBDD de HIVE y creamos una tabla con desde la línea de comandos el siguiente código:

CREATE TABLE Top\_12\_German\_Companies (

Company STRING,

Period STRING,

Revenue DOUBLE,

Net\_Income DOUBLE,

Liabilities DOUBLE,

Assets DOUBLE,

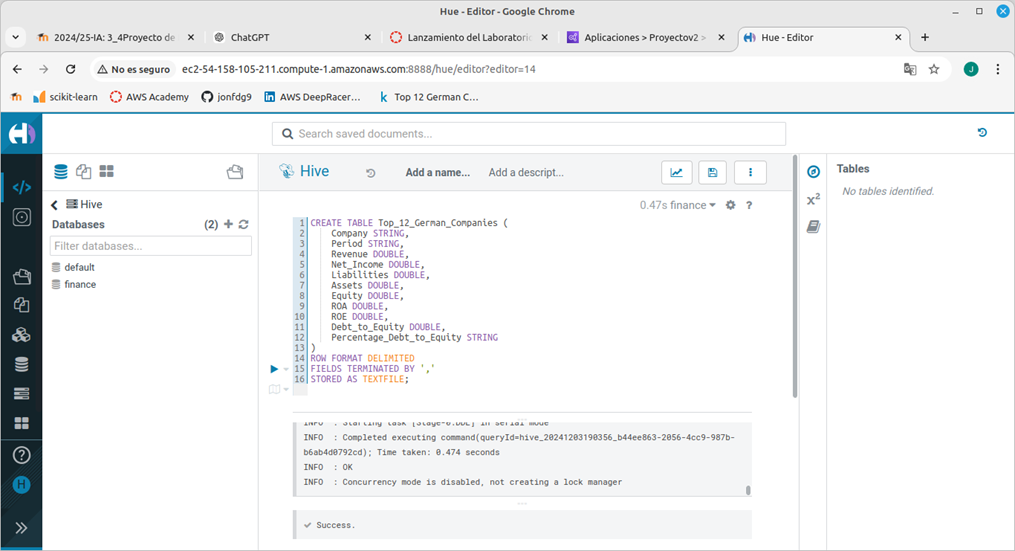
Equity DOUBLE

)

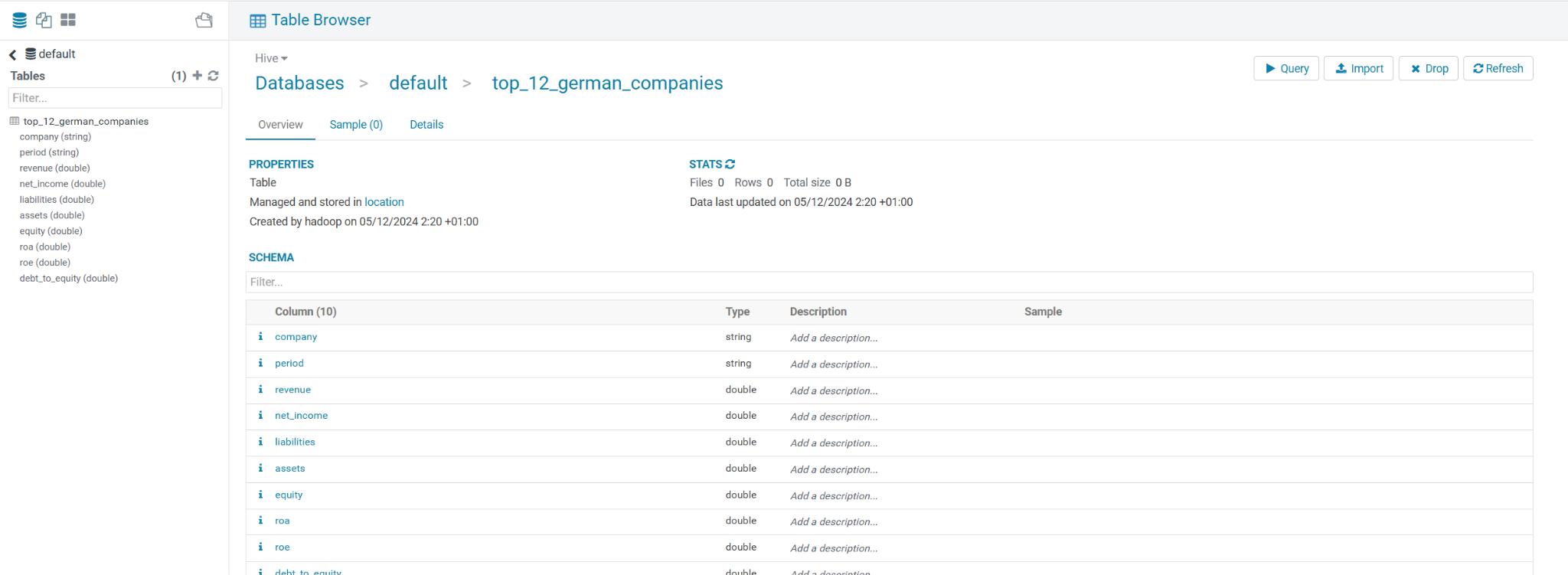
ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;



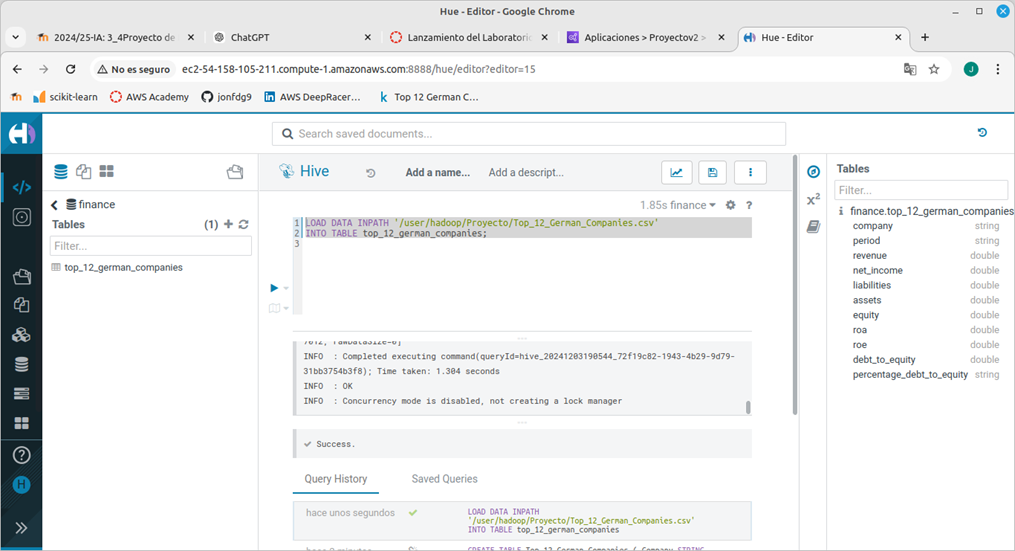
Una vez creada la tabla comprobamos que nos ha añadido todas las columnas y los tipos de datos de cada una correctamente.



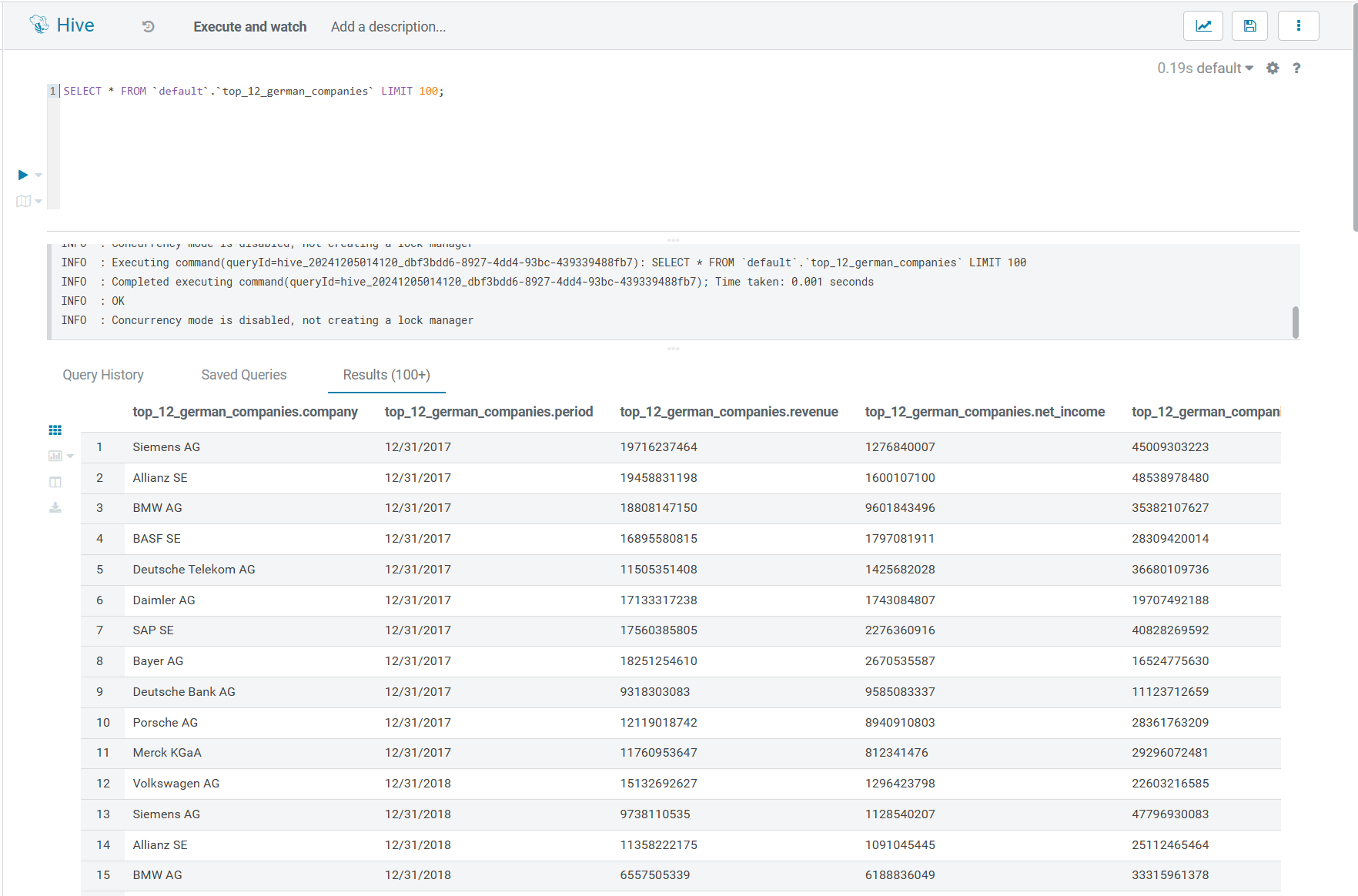
Ahora tenemos que cargar los datos que tenemos dentro del hdfs.

LOAD DATA INPATH'/user/hadoop/Proyecto/Top\_12\_German\_Companies.csv'

INTO TABLE top\_12\_german\_companies;



Vemos que los datos insertados son correctos



## CRUD

### Create

Al crear la BBDD lo hemos hecho, las capturas están arriba.

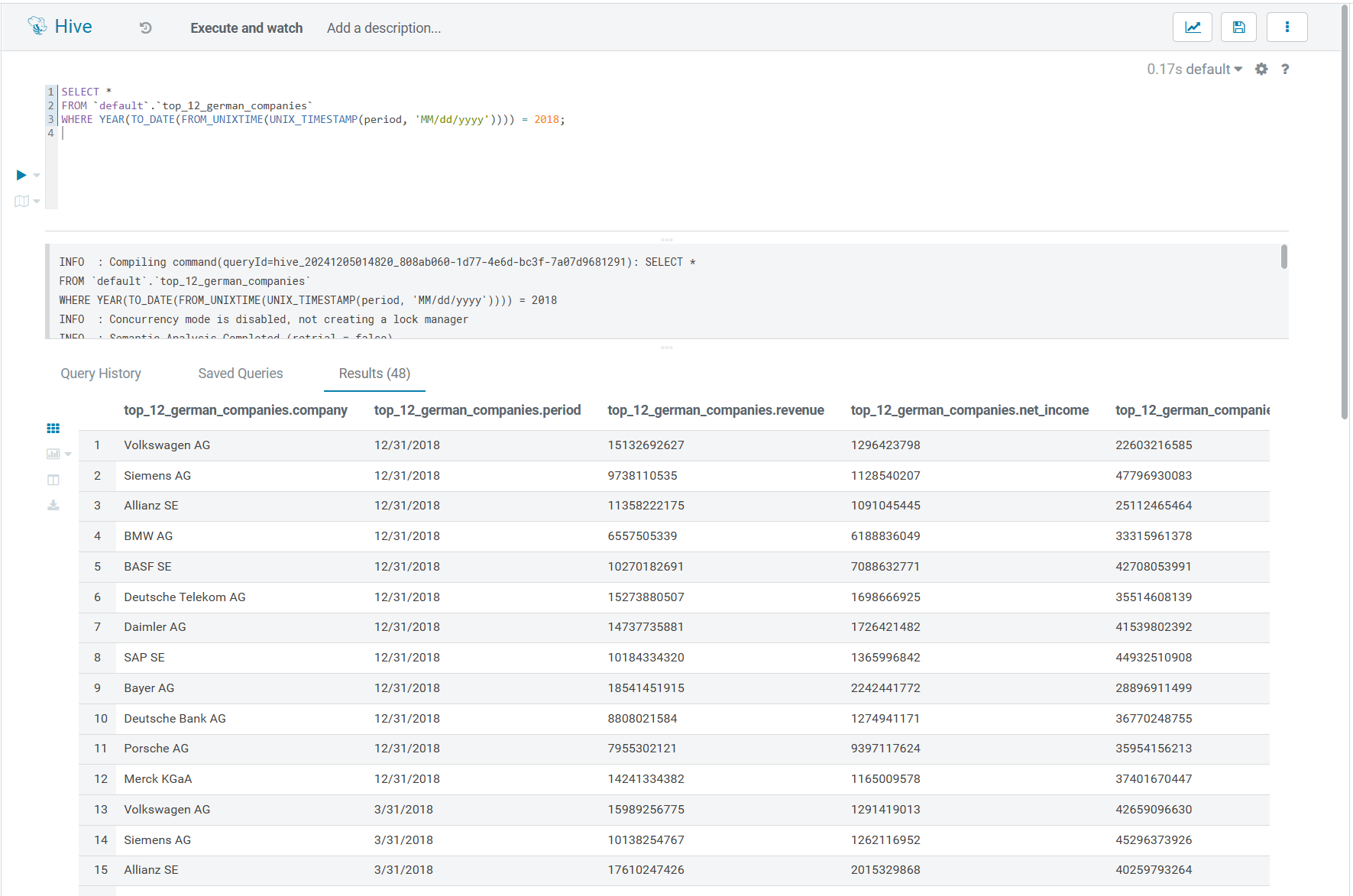
### Read

Hacemos una consulta que nos devuelva todos los datos del 2018, tenemos que hacer la transformación de la fecha porque en nuestra BBDD está como String y no como date.

SELECT \*

FROM `default`.`top\_12\_german\_companies`

WHERE YEAR(TO\_DATE(FROM\_UNIXTIME(UNIX\_TIMESTAMP(period, 'MM/dd/yyyy')))) = 2018;

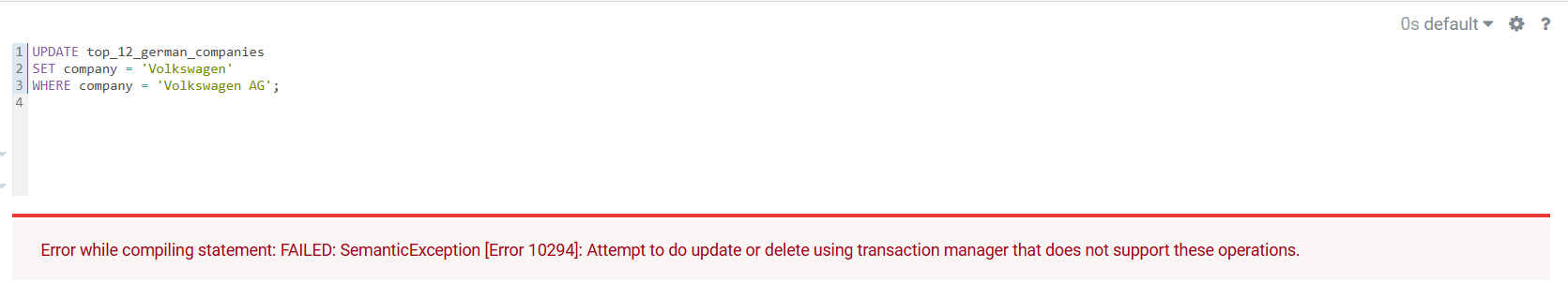


### Update

UPDATE top\_12\_german\_companies

SET company = 'Volkswagen'

WHERE company = 'Volkswagen AG';

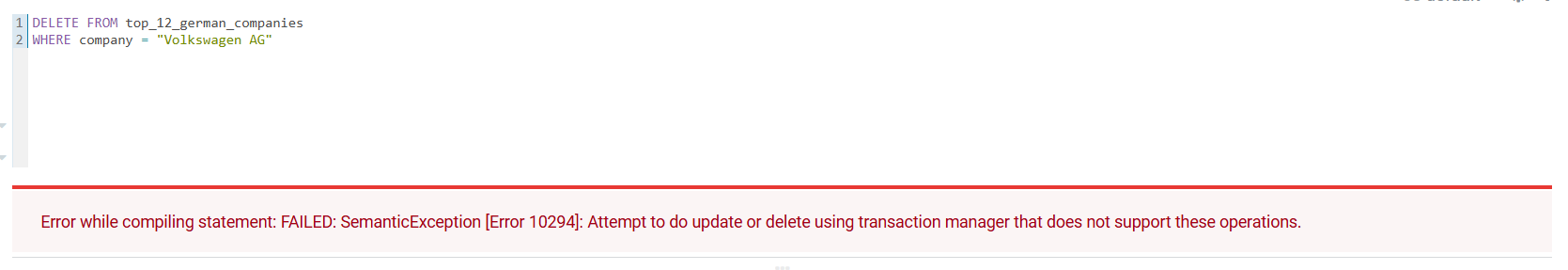


Al intentar hacer un update nos produce un error, aunque la consulta sea correcta.

### Delete

DELETE FROM top\_12\_german\_companies

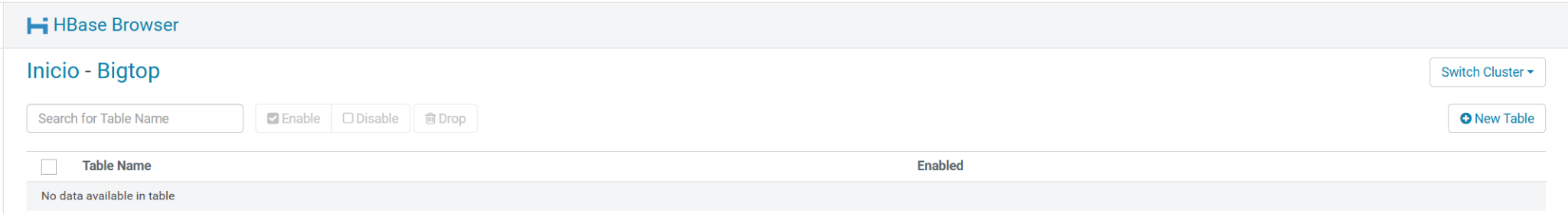
WHERE company = "Volkswagen AG"



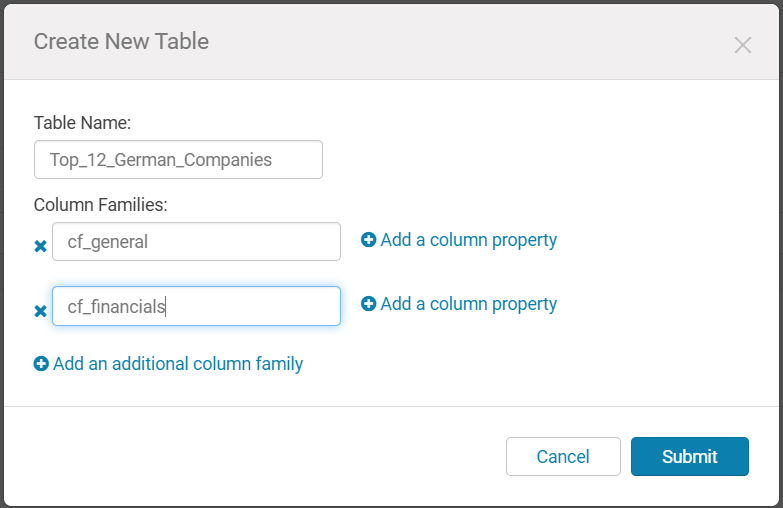
Al igual que con las modificaciones, tampoco nos deja eliminar filas.

# HBASE

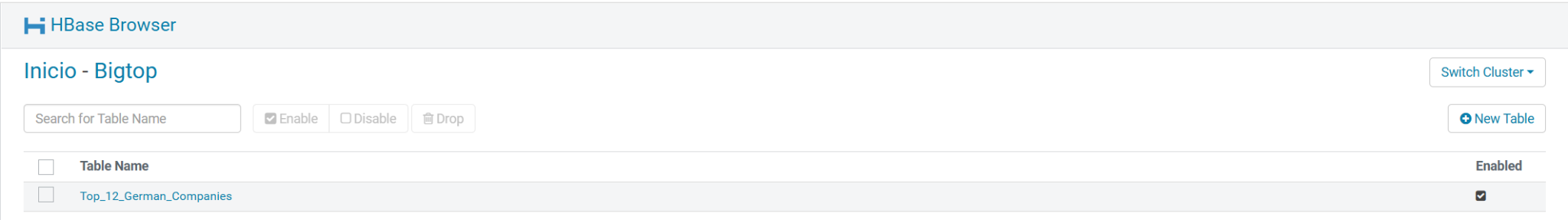
Dentro de HUE nos dirigimos a la pestaña de HBase y lo primero que tenemos que hacer es crear una nueva tabla.



Dentro de la creación le indicamos cuantas familias de columnas le queremos añadir.



Ya podemos ver que nos la ha creado.



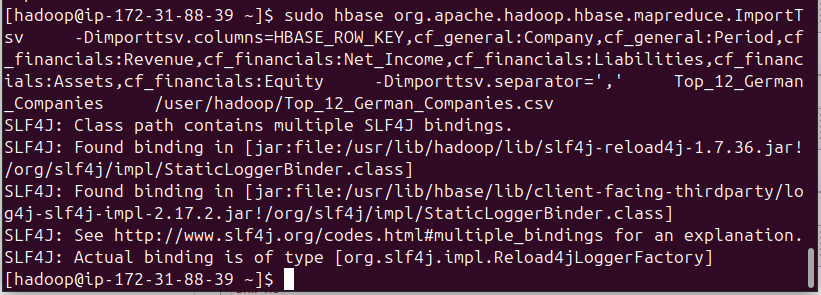
Ahora nos vamos a la línea de comandos, al entorno donde tenemos instalado el hadoop y cargamos el csv que tenemos en hdfs a la tabla que hemos creado de HBase.

sudo hbase org.apache.hadoop.hbase.mapreduce.ImportTsv \ -Dimporttsv.columns=HBASE\_ROW\_KEY,cf\_general:Company,cf\_general:Period,cf\_financials:Revenue,cf\_financials:Net\_Income,cf\_financials:Liabilities,cf\_financials:Assets,cf\_financials:Equity \

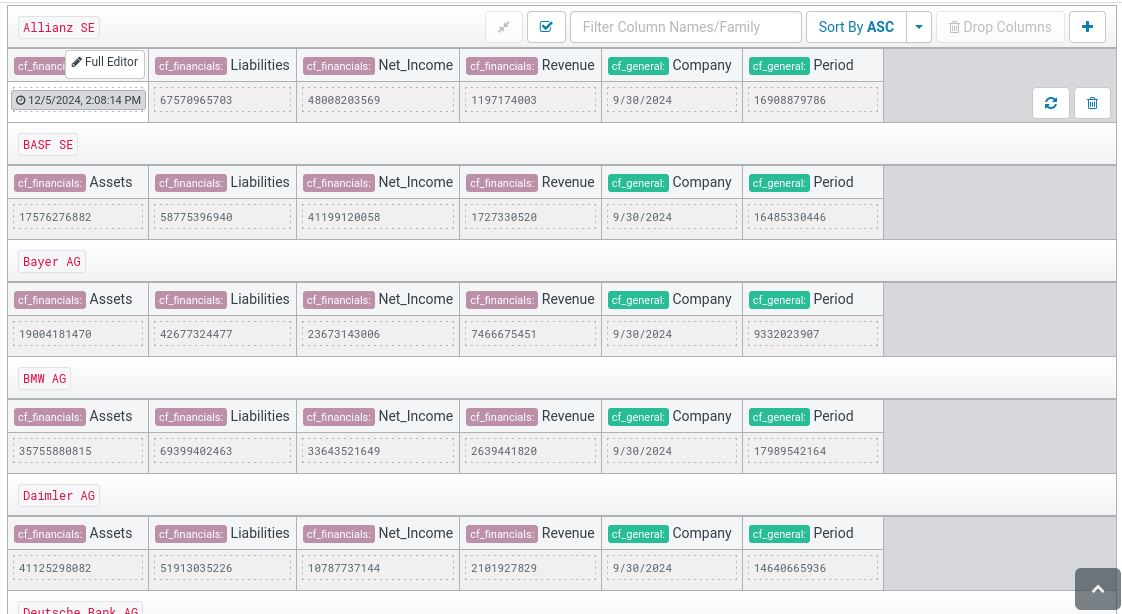
-Dimporttsv.separator=',' \

Top\_12\_German\_Companies \

/user/hadoop/Top\_12\_German\_Companies.csv



cf\_financials van a contener los datos Revenue, Net\_Income, Liabilities, Assets y Equity, mientras que la columna cf\_general va contener información adicional como Company y Period.



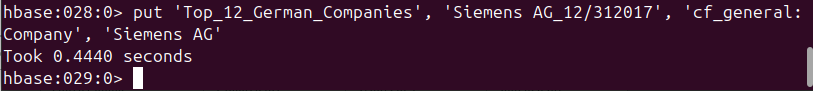
Ahora nos vamos a conectar a la shell de HBase para hacer las consultas del CRUD.



## CRUD

### Create

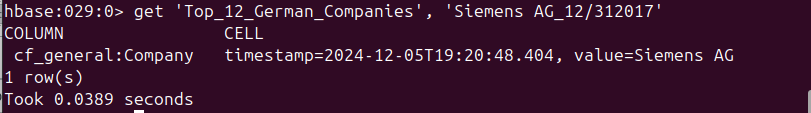
put 'Top\_12\_German\_Companies', 'Siemens AG\_12/31/2017', 'cf\_general:Company', 'Siemens AG'



### 

### Read

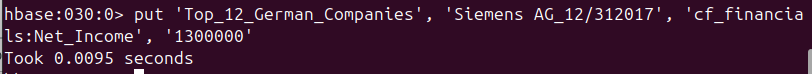
get 'Top\_12\_German\_Companies', 'Siemens AG\_12/31/2017'



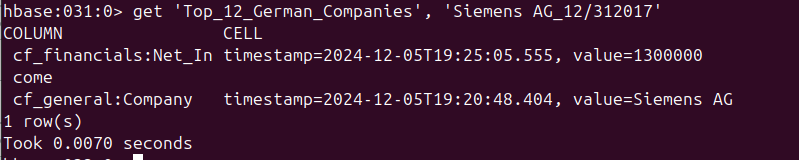
### Update

En hbase el comando put crea en caso de que no exista pero también modifica en el caso en el que si lo haga.

put 'Top\_12\_German\_Companies', 'Siemens AG\_12/31/2017', 'cf\_financials:Net\_Income', '1300000000'



Comprobación

****

### Delete

deleteall 'Top\_12\_German\_Companies', 'Siemens AG\_12/31/2017'

Para eliminar toda la fila con la clave Siemens AG\_12/31/2017:

