# PROYECTO 1

Hemos visto que Hadoop es un ecosistema que consta de diferentes herramientas para almacenar, procesar y analizar datos Big Data (Petabytes: 1015 byte).

Hasta ahora hemos aprendido cómo instalar un clúster Hadoop en máquinas virtuales, así como en AWS y cómo lanzar tareas utilizando el motor MapReduce.

Para avanzar en el estudio de las herramientas del ecosistema Hadoop en este proyecto nos centraremos en Hive (SQL), HBase(NoSQL) y HUE (interfaz gráfica).

Lo que vamos a hacer es descargarnos desde Kaggle algunos ficheros, subirlos a Hadoop y utilizando Hive y HBase haremos consultas sobre sus datos, la herramienta HUE la podemos usar como interfaz.

Los pasos a seguir son:

1.- Crear un clúster en AWS con las siguientes aplicaciones: Hadoop, HBase, Hive y Hue.

2.- Descargar un repositorio de Kaggle.

3.- Subir los archivos Hadoop (AWS)

4.- Crear un pequeño manual de las instancias CRUD HBase y Hive.

5.- Utilizando Hive y HBase realizar consultas acerca de estos archivos subidos.

5.1.- Estudiar los datos almacenados en los archivos para crear las tablas en las que cargar los datos.

5.2.- Crear las tablas en HBase y Hive.

5.3.- Cargar los datos de los ficheros en las tablas.

5.4.- Crear diferentes consultas sobre las tablas y mostrar el resultado obtenido.

Todo (salvo los pasos para crear el clúster Hadoop que se da por hecho que ya lo conocemos) tiene que estar documentado con ejemplos ejecutados en vuestro propio equipo o máquina virtual:

* Código a ejecutar en formato texto (para poder cortar y pegar)
* Pantallazo del código ejecutado

## CRITERIOS DE EVALUACIÓN

* (10p) Documentación:
  + Portada (0.4p)
  + Índice con número de página (enlaces) (0.4p)
  + Número de página(0.4p)
  + Buena presentación (alineaciones, espacios en blanco innecesarios, imágenes no colocadas correctamente) (0.4p)
  + Faltas de ortografía (0.4p)
  + En cuanto a su contenido se valorará: (8p)

Si contiene todos los puntos y se explican los conceptos para que los entienda cualquier persona que lea la documentación. (6p)

El lenguaje utilizado: Si se utilizan los términos técnicos adecuadamente.(2p)

La nota de este proyecto tiene un peso del 30% sobre la nota del módulo.

CREATE TABLE Top\_12\_German\_Companies (

Company STRING,

Period STRING,

Revenue DOUBLE,

Net\_Income DOUBLE,

Liabilities DOUBLE,

Assets DOUBLE,

Equity DOUBLE,

ROA DOUBLE,

ROE DOUBLE,

Debt\_to\_Equity DOUBLE,

Percentage\_Debt\_to\_Equity STRING

)

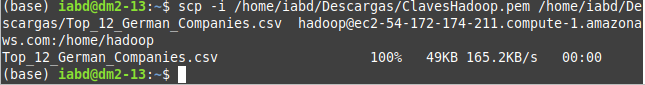
ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;

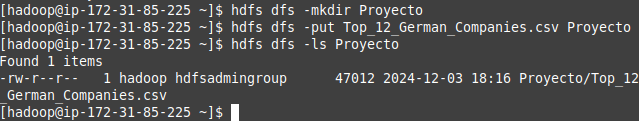
Subimos el archivo csv que hemos descargado de Kaggle al entorno de hadoop

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

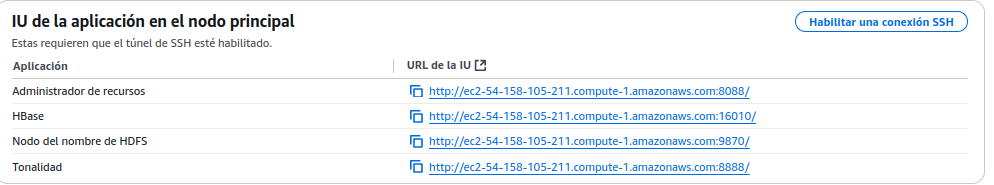


Tenemos que subirlo ahora del entorno al gestor de archivos de hadoop

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



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

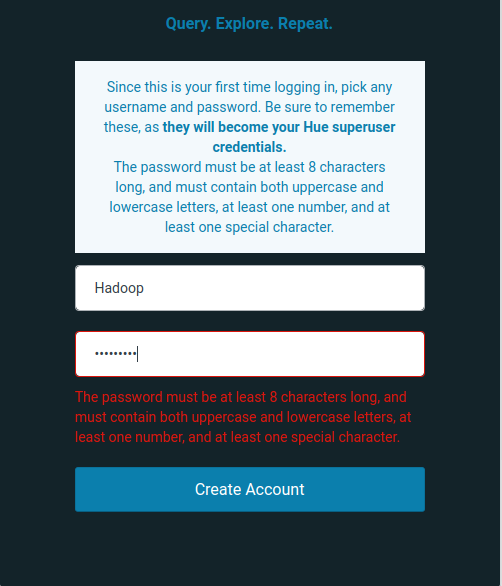


Nos pide crear un usuario

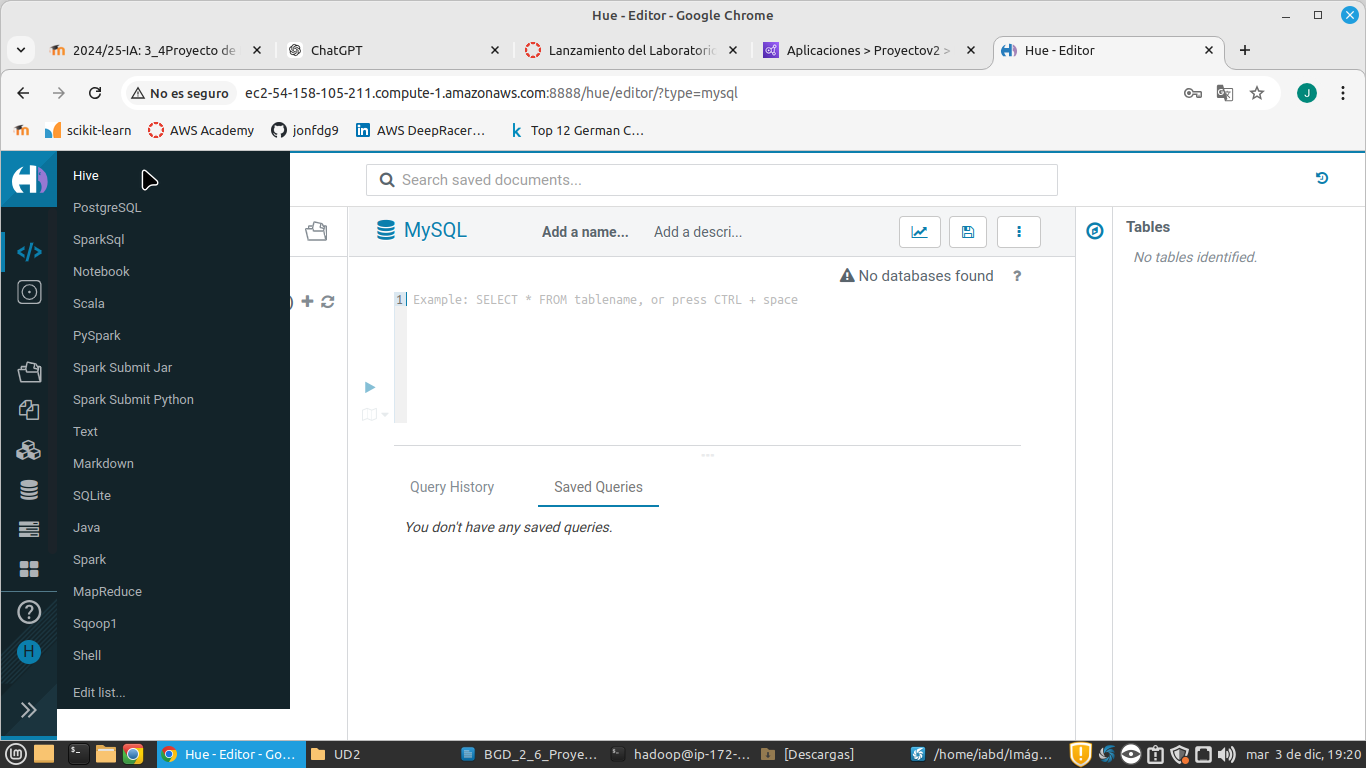
HUE

User: Hadoop

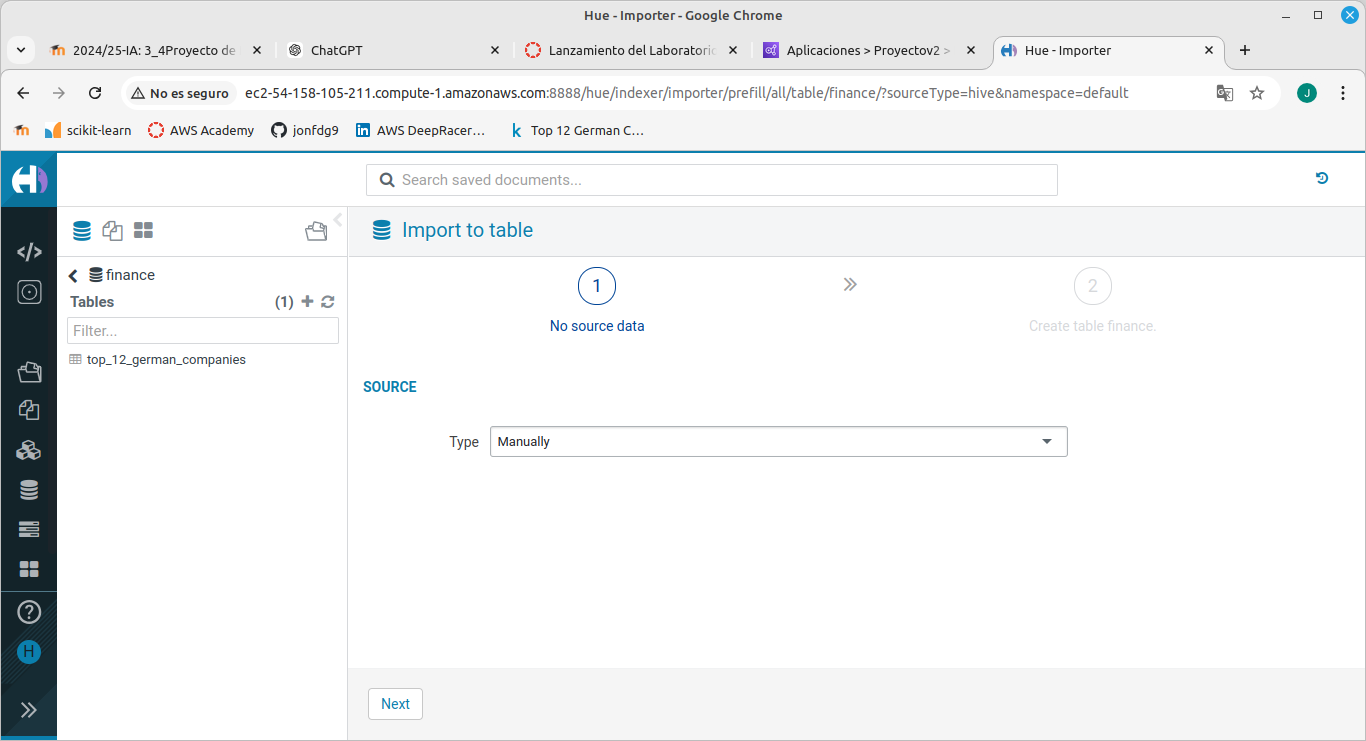
Pass: Hadoop1@



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



Creamos una BBDD y dentro de ella una tabla



CREATE TABLE Top\_12\_German\_Companies (

Company STRING,

Period STRING,

Revenue DOUBLE,

Net\_Income DOUBLE,

Liabilities DOUBLE,

Assets DOUBLE,

Equity DOUBLE,

ROA DOUBLE,

ROE DOUBLE,

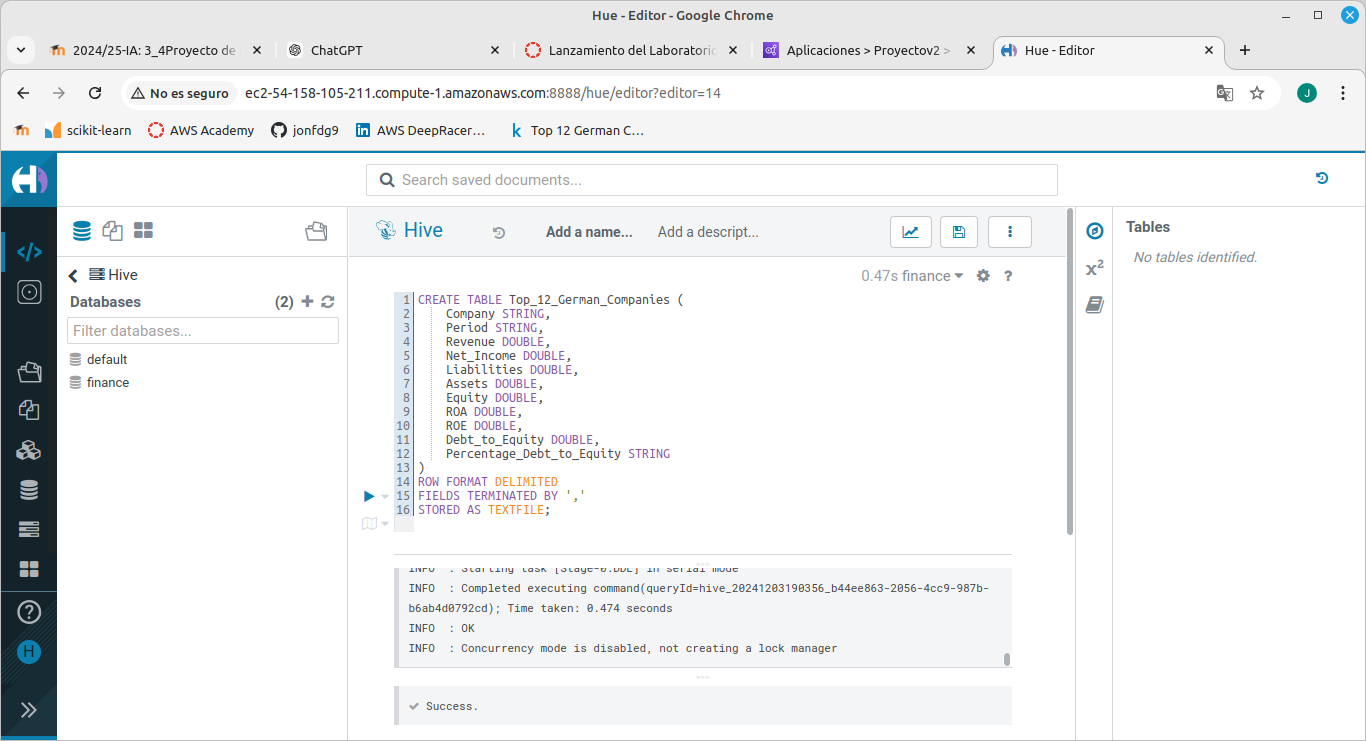
Debt\_to\_Equity DOUBLE

)

ROW FORMAT DELIMITED

FIELDS TERMINATED BY ','

STORED AS TEXTFILE;



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

INTO TABLE top\_12\_german\_companies;

