**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**

**Proyecto Final**

**Sistema de Bases de Datos Avanzadas**

# **Cubo de información con dataset Adventure works**

**Grupo 5**

**INTEGRANTES**

**Alex Velez**

**Valeria Barzola**

## **Término**

## **2021-2S**

## **Objetivos del proyecto**

Creación de un cubo de datos de una datawarehouse de Adventure Works para la extracción de insights que ayuden a la toma de decisiones del negocio.

## **Descripción del datase**t

Adventure Works es una empresa que vende bicicletas y partes de bicicletas a clientes y distribuidores. En la base de datos provista tenemos siguientes datos como:

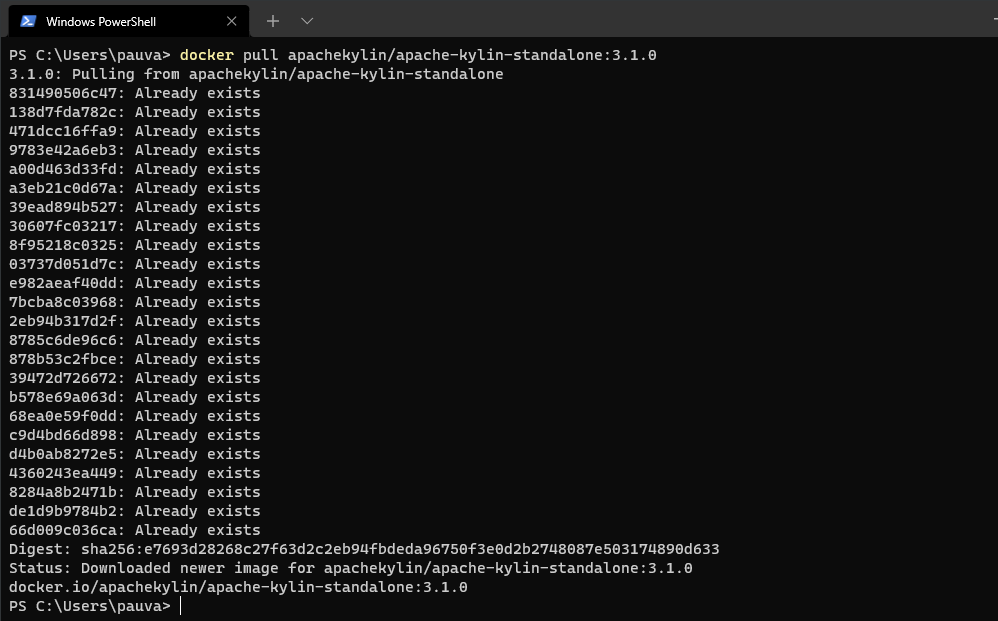
* Clientes
* Productos (cateogorías, subcatecogrías)
* Ventas
* Territorio de ventas

**Instalación del apache kylin**

Apache Kylin es una Data Warehouse distribuida, utilizada para el Big Data. Para instalarla con docker debemos seguir los siguientes pasos. La instalación se hará en una máquina virutal de Linux (ubuntu 20.04).

1. Hacerle pull a la imagen oficial de Kylin version 3.1.0 con el siguiente comando:

docker pull apachekylin/apache-kylin-standalone:3.1.0



1. Correr el contenedor con el siguiente comando:

docker run -d \

-m 8G \

-p 7070:7070 \

-p 8088:8088 \

-p 50070:50070 \

-p 8032:8032 \

-p 8042:8042 \

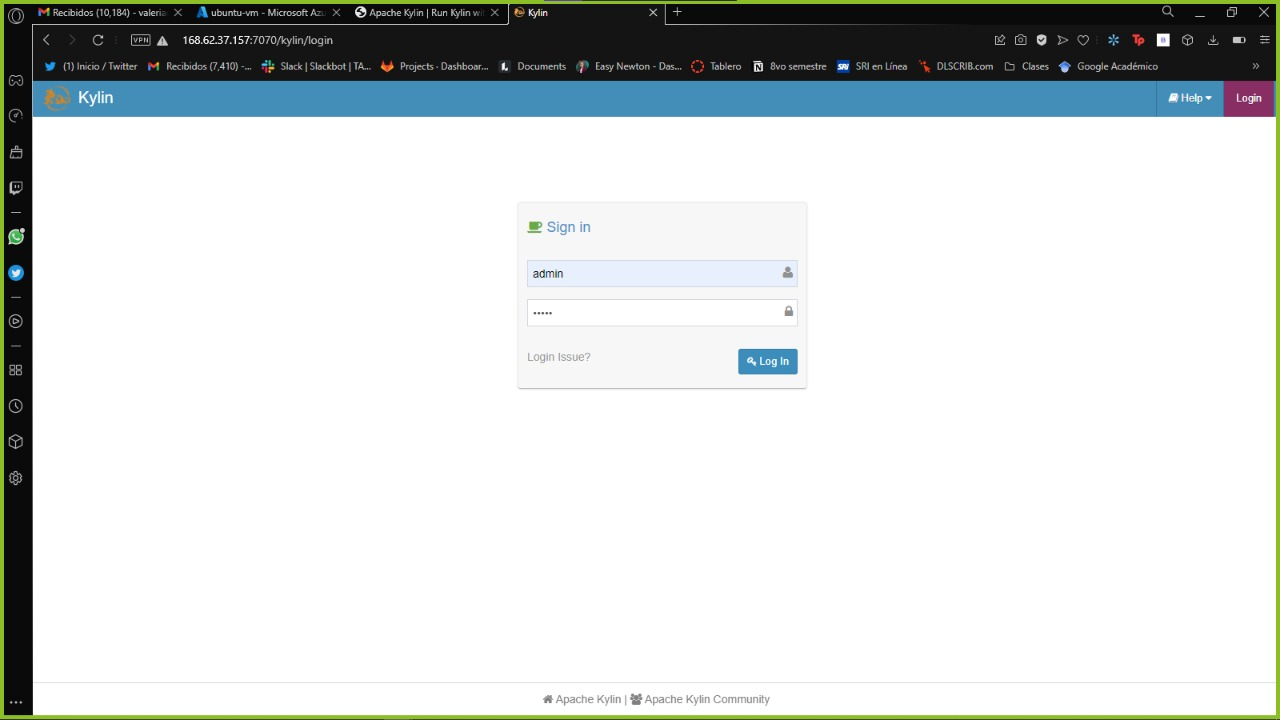
-p 16010:16010 \

apachekylin/apache-kylin-standalone:3.1.0

1. Con el id obtenido, executar la interfaz interactiva del contendor con el siguiente comando:

docker exec -it <container\_id> bash

1. Con todo esto podremos ingresar a la interfaz de Kylin a través del puerto 7070



## **Importación de datos**

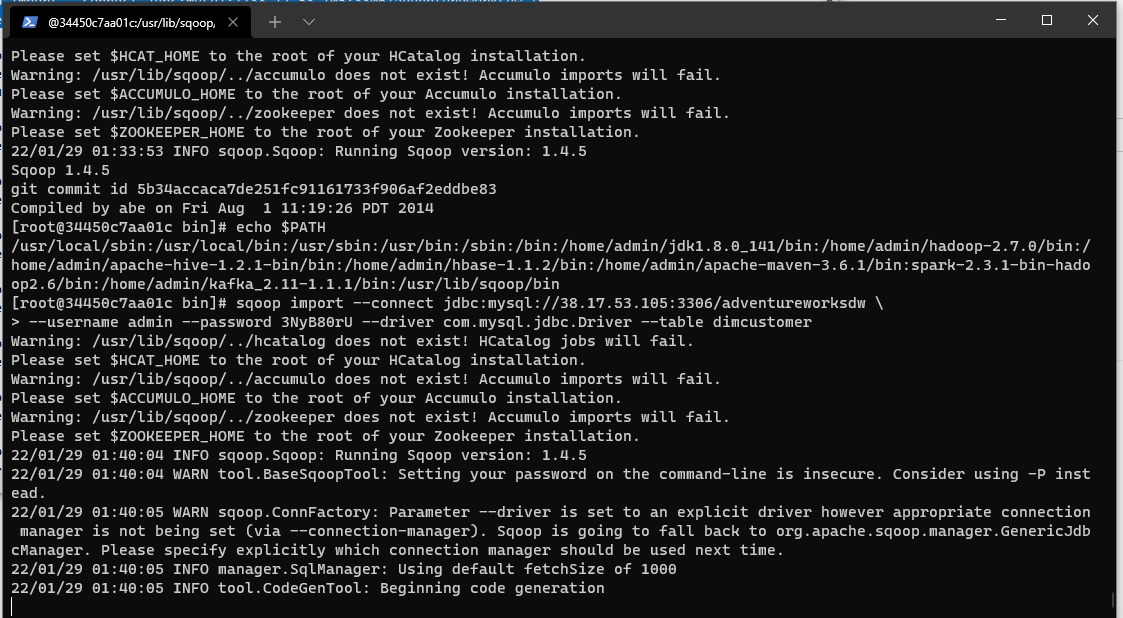
Para importar los datos desde la base de datos mysql hacia el sistema de archivos de hadoop, podemos utilizar la herramienta sqoop. Una vez instalado, podemos ejecutar el siguiente comando para importar los datos:

sqoop import --connect jdbc:mysql://<DATABASE\_IP>:<DATABASE\_PORT>/<DATABASE\_NAME> --username <DATABASE\_USERNAME> --password <USERNAME\_PASSWORD> --table <TABLE\_TO\_IMPORT>

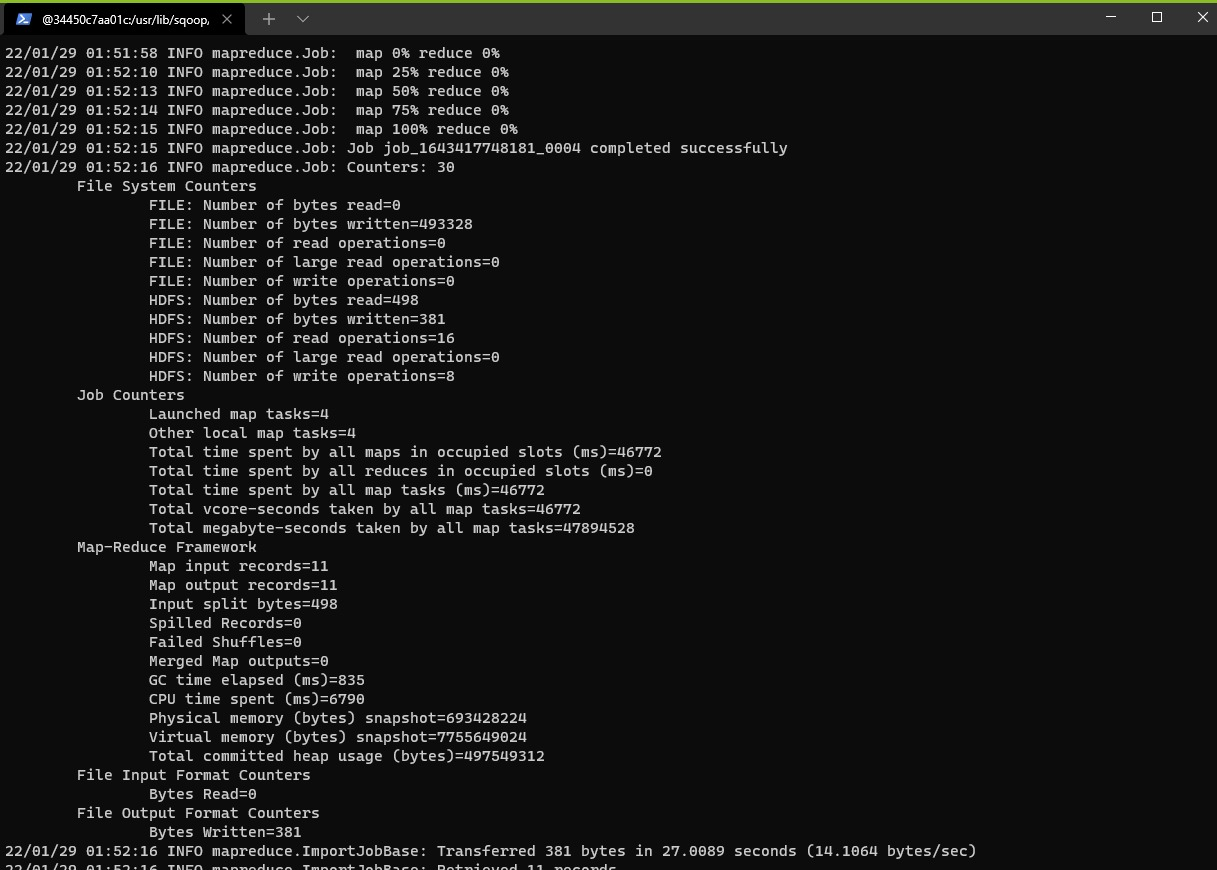
Donde:

* <DATABASE\_IP>: Representa la ip de la base de datos donde está almacenada la imformación que deseamos cargar.
* <DATABASE\_PORT>: Representa el puerto de la base de datos donde está almacenada la información que deseamos cargar.
* <DATABASE\_NAME>: Representa el nombre de la base de datos donde está almacenada la información que deseamos cargar. En nuestro caso, la base tiene el nombre de **adventureworksdw**.
* <DATABASE\_USERNAME>: Representa el nombre de usuario con el que se va a acceder a la base para cargar la información.
* <USERNAME\_PASSWORD>: Representa la contraseña del usuario con el que se va a acceder a la base para cargar la información.
* <TABLE\_TO\_IMPORT>: Representa el nombre de la tabla donde está almacenada la información que deseamos cargar. En general iremos corriendo varias veces este comando por cada tabla a importar.

Por ejemplo, si deseamos importar los datos de la tabla **dimcustomer** hacia el sistema de archivos de hadoop la ejecución sería la siguiente:



Este va a ir importando los datos desde mysql al hdfs, realizando map reduces hasta que finalmente veremos algo como lo siguiente:



Donde podemos observar información acerca del job de map reduce, y cuántos bytes se han transferido.

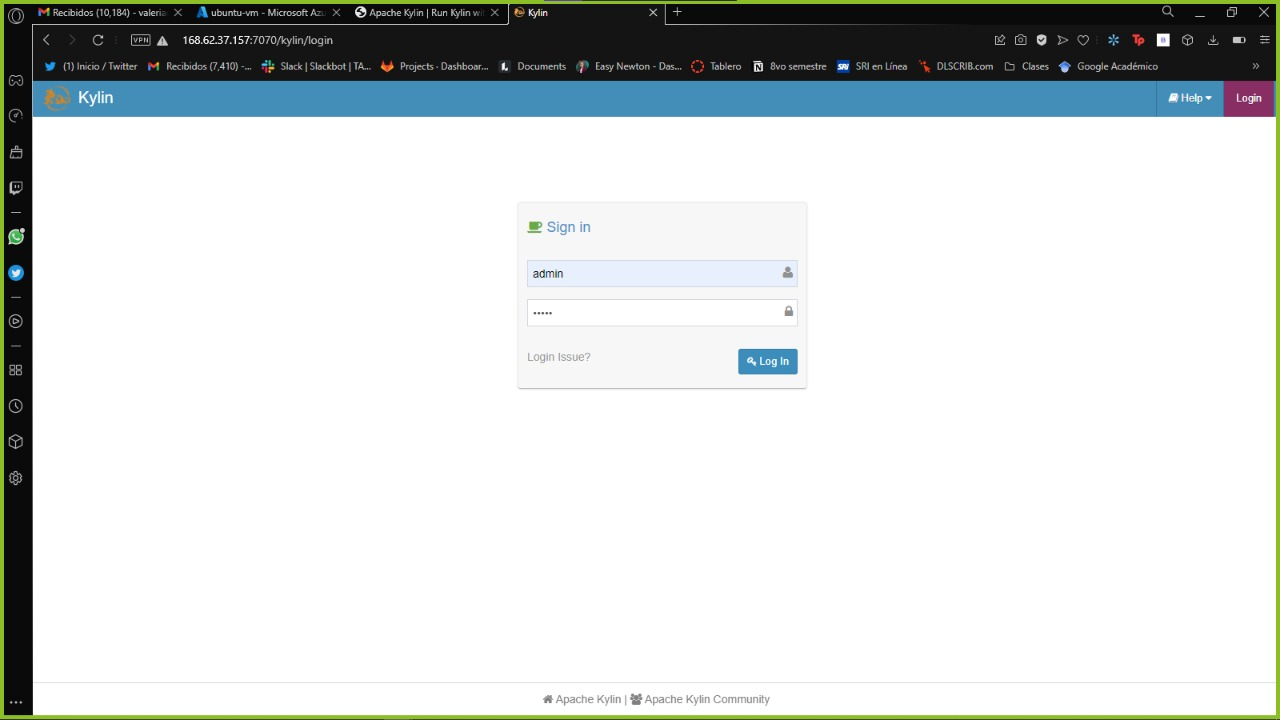
Tenemos que correr este comando con todas las tablas que deseamos importar, en nuestro caso fueron las siguientes:

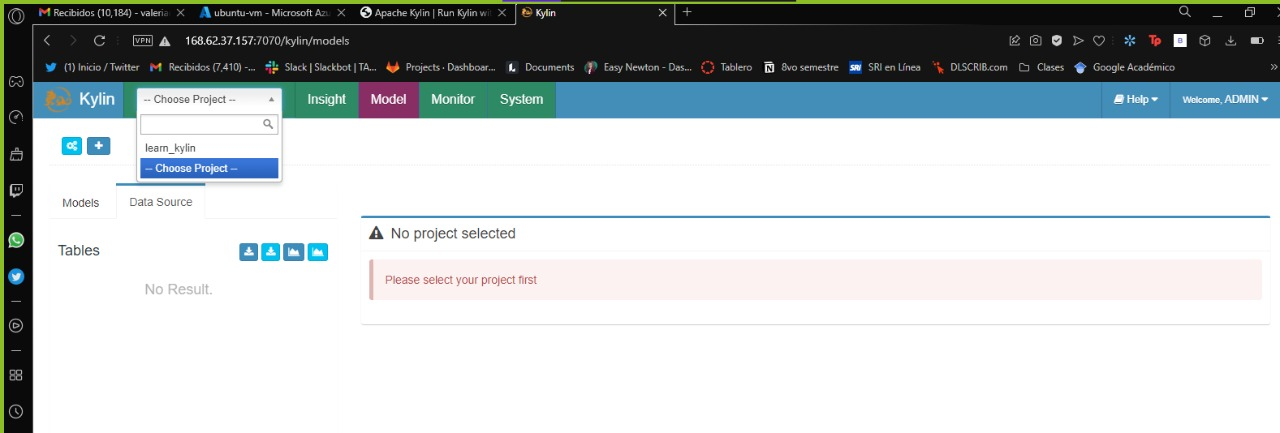
* DimCurrency
* DimCustomer
* DimDate
* DimGeography
* DimProduct
* DimProductCategory
* DimProductSubcategory
* DimSalesTerritory
* FactInternetSales

Que corresponden a las 8 dimensiones y la tabla de hechos que se usará para el presente proyecto.

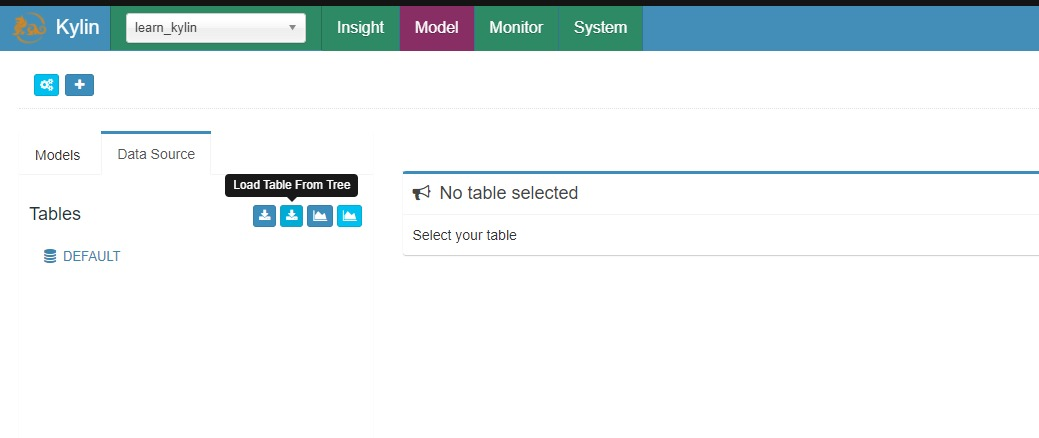
## **Creación del modelo**

El modelo será creado con ayuda de la interfaz de Kylin, siguiendo lo siguientes pasos:

1. En el navegador visitar el puerto **7070** para ingresar a la interfaz de kylin, aquí nos autenticamos utilizando el usuario **admin** y la contraseña **KYLIN.** ****
2. Seleccionar el proyecto haciendo click en el select que dice **Choose Project**



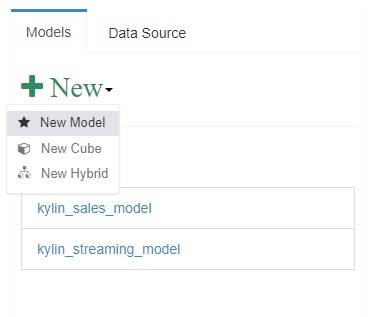
1. En la pestaña **Data source** de la izquierda, seleccionar “load table from tree”



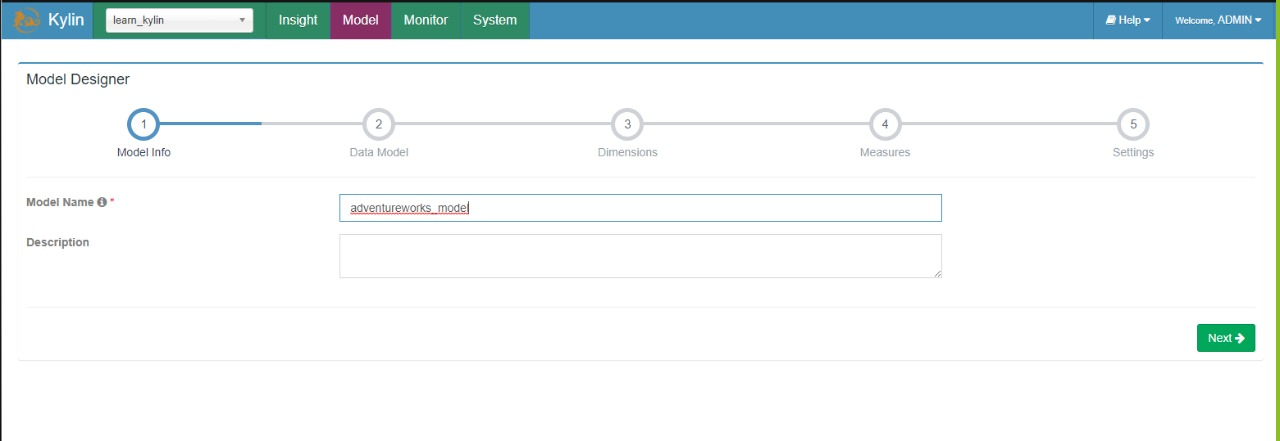
1. Aquí nos aparecerá la base de datos que importamos previamente. Seleccionamos todas las tablas y hacemos click en **sync**

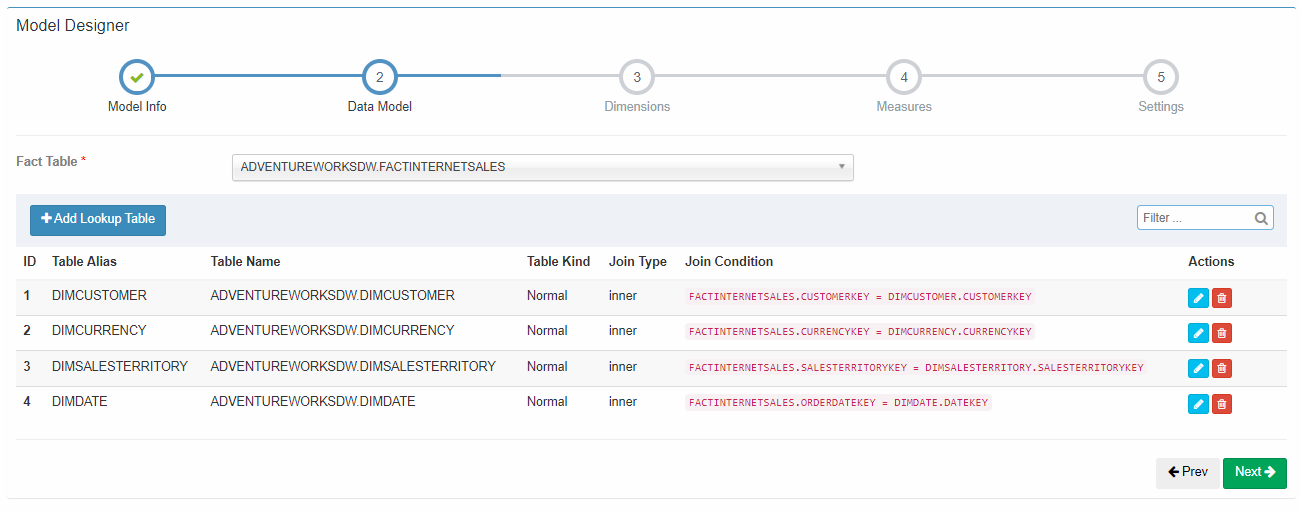


1. Ahora sí procederemos a crear el modelo, haciendo click en **NEW** y luego en **new Model**



1. Como primer paso de la creación del modelo, le asignaremos un nombre al mismo. Luego hacemos click en **next.**



1. En el segundo paso, escogeremos cuales será la tabla del hecho que queremos modelar, en este caso FACTINTERNETSALES. Además seleccionares como se relacionará el hecho con cada dimensión, haciendo click en el botón de **Add lookup table.**
2. Elegimos los campos necesarios de cada dimensión.

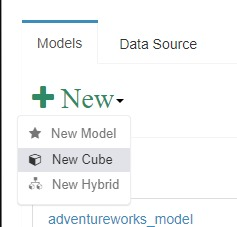
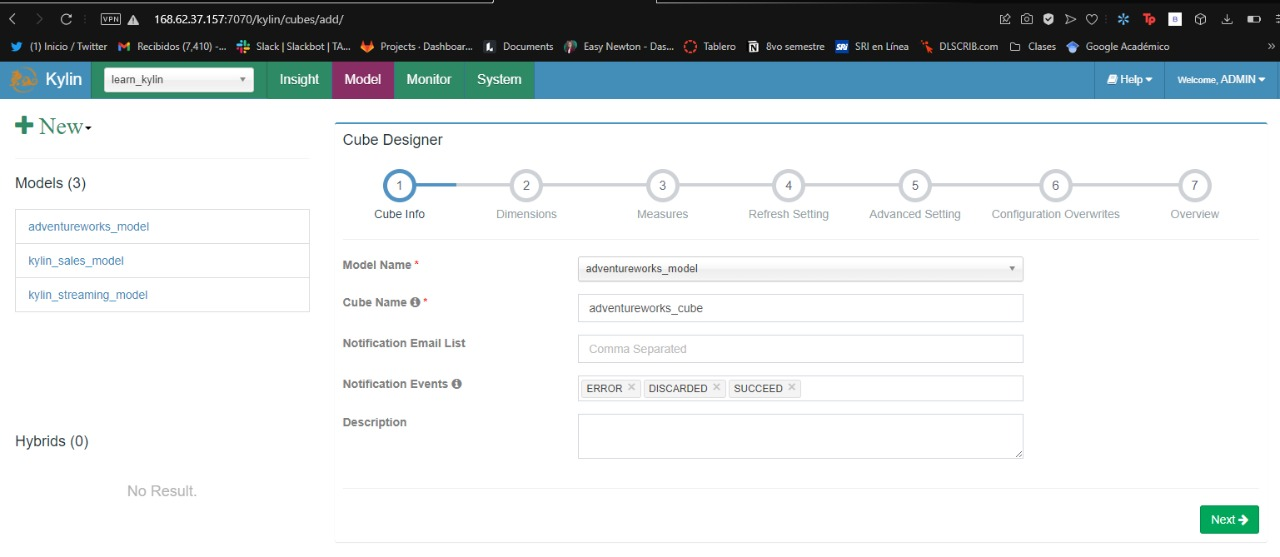
## 

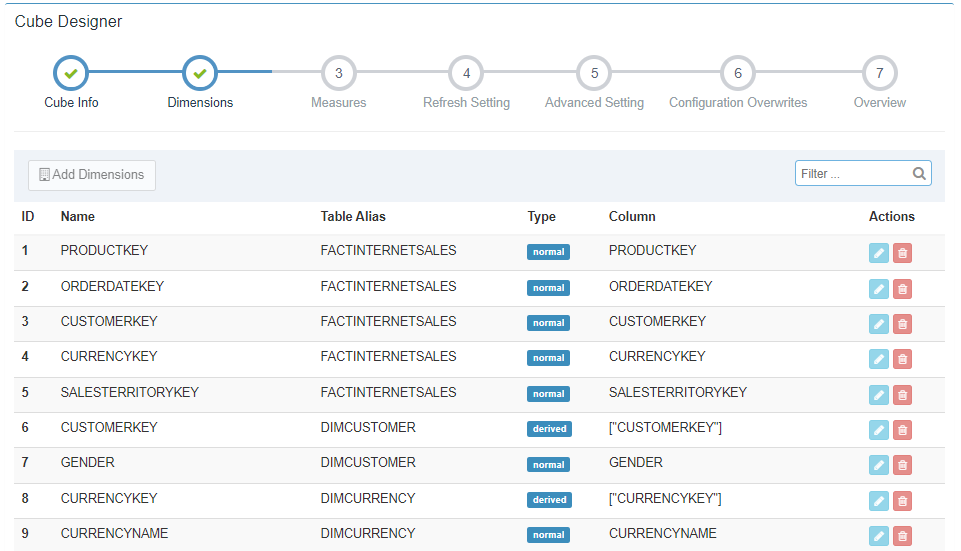
1. Finalmente seleccionamos los campos necesarios del Hecho y listo, el modelo ha sido creado

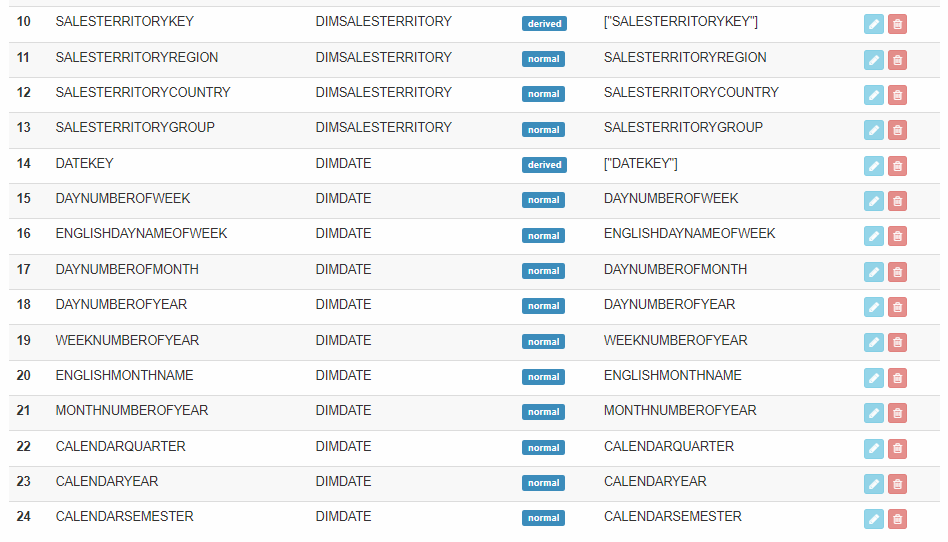
## 

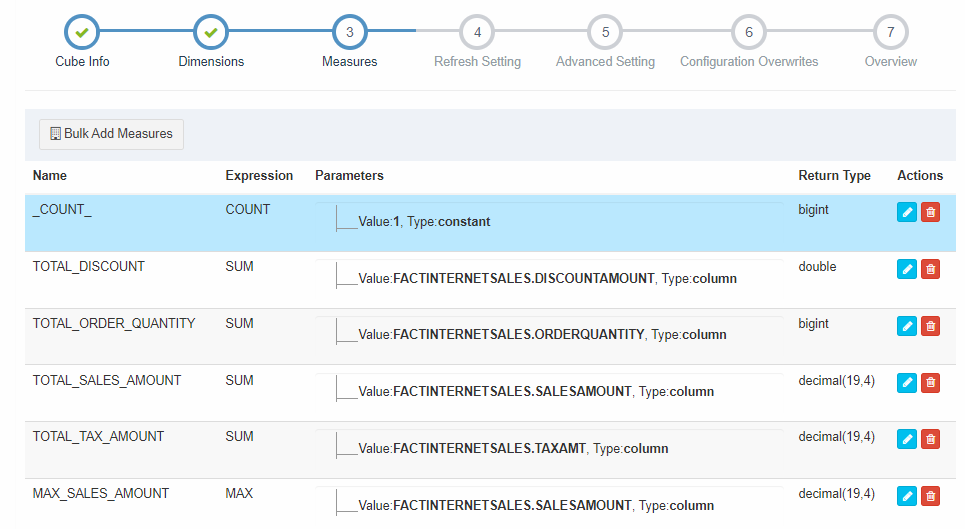
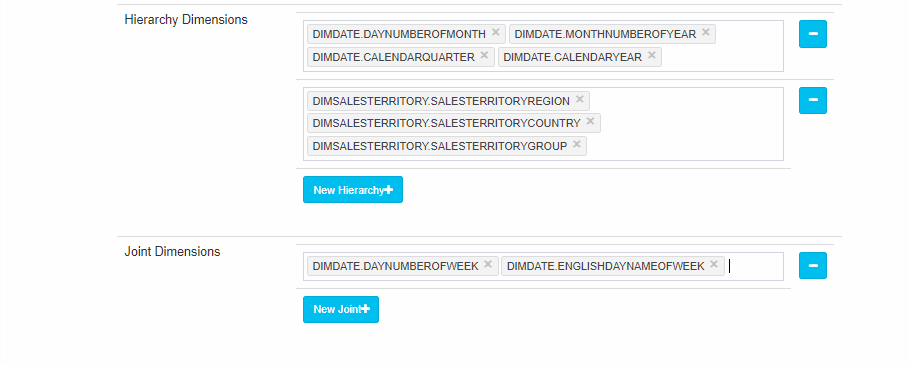
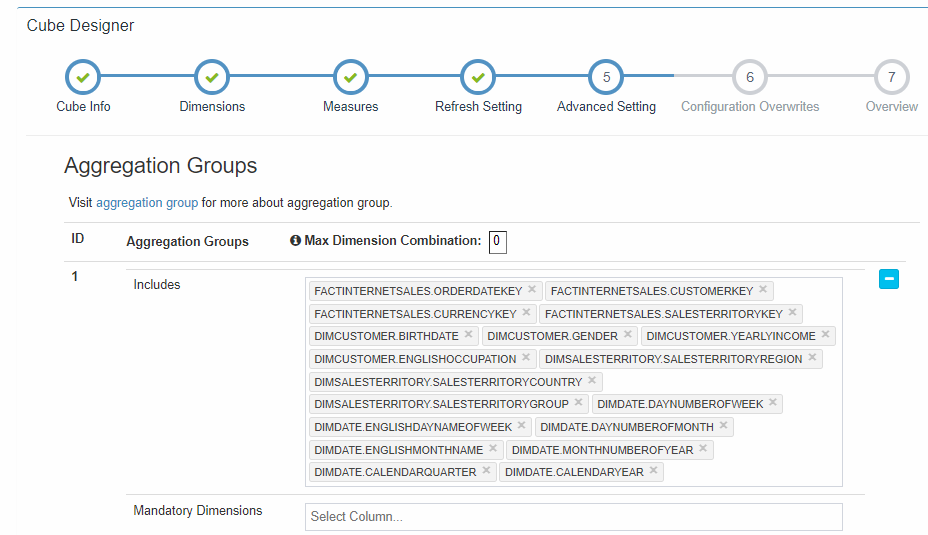
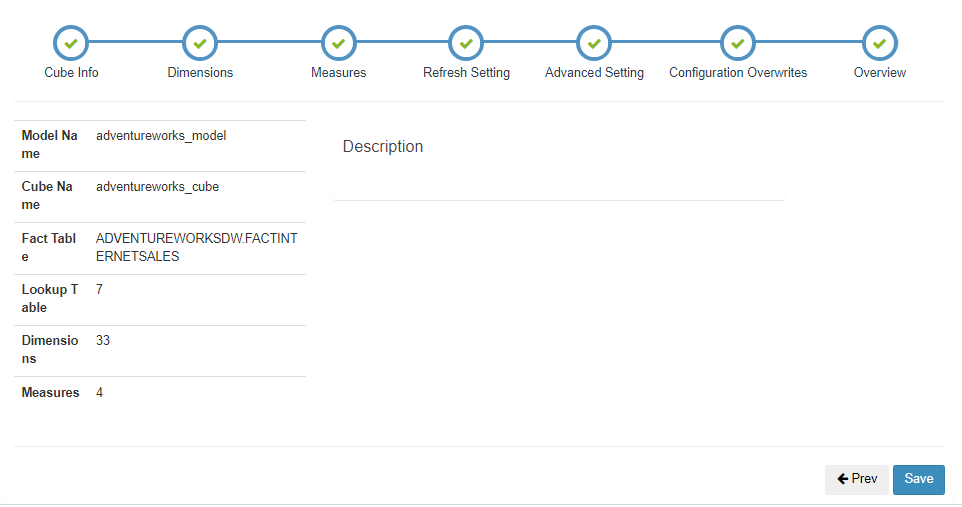
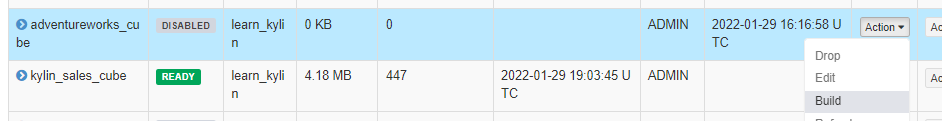
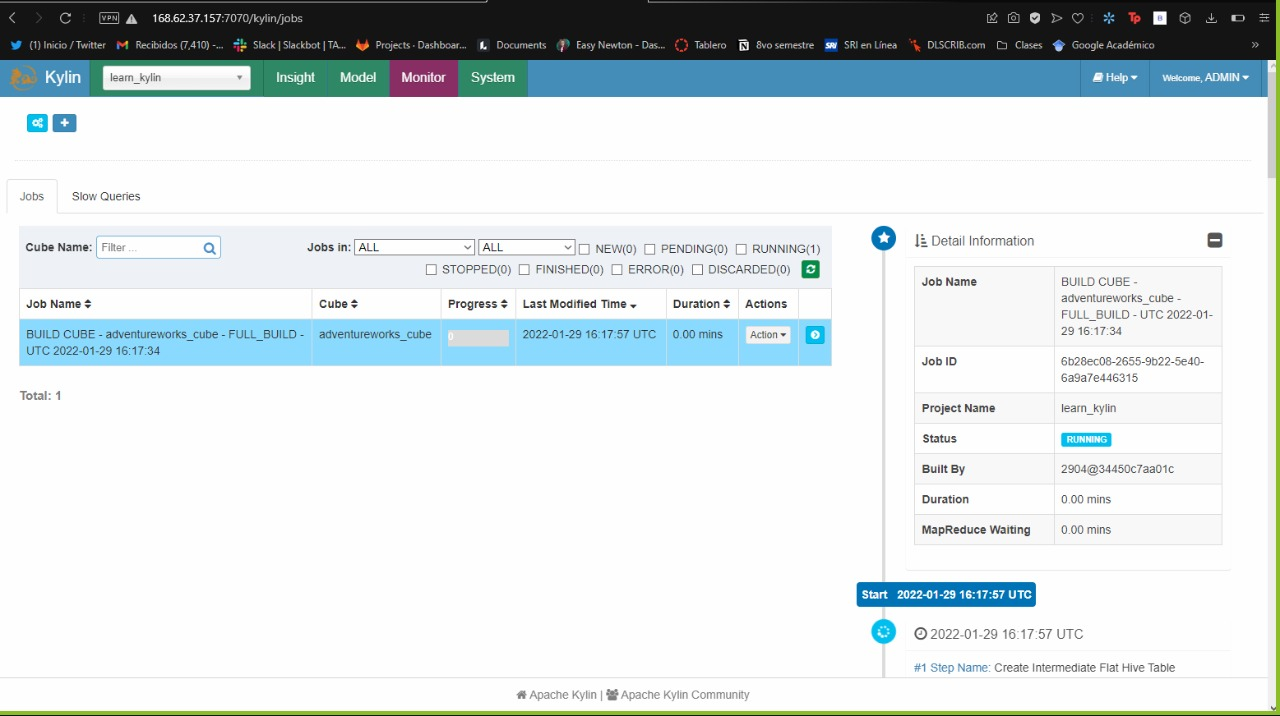
## **Creación del cubo**

Una vez creado el modelo, podemos continuar con la creación del cubo siguiendo los siguientes pasos:

1. En la pestaña Model de la izquierda, damos click en **New** y luego en **New cube.**
2. Luego, nos aparecerá la interfaz de **Cube designer** para diseñar nuestro cubo. En el primer tab pondremos la información general del cubo, como su nombre, y elegiremos el modelo que deseamos utilizar, para ello escogemos el modelo creado previamente. Podemos además agregar información para que nos notifique sobre eventos, pero en este caso lo dejaremos por defecto.
3. En el segundo tab pondremos la información acerca de las dimensiones de nuestro cubo, para ello, utilizaremos las dimensiones mostradas en la siguiente imagen. Además, por cada dimensión se especifica su condición de Join, en la que simplemente se conectarán los foreign keys de la tabla de hechos, con las primary keys de las demás dimensiones.





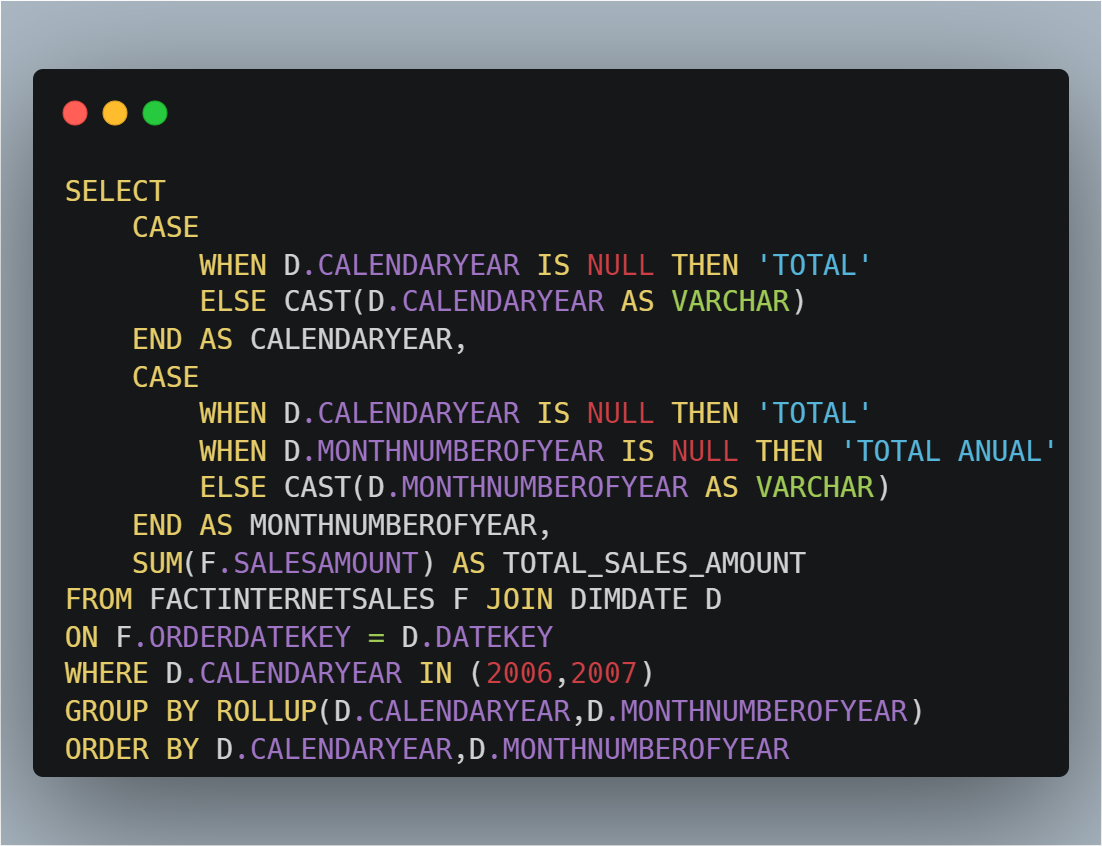
1. En la siguiente pestaña, indicaremos cuáles son los measures que queremos obtener, y su función de sumarización, en nuestro caso obtendremos el total de la cantidad de ventas a través de la suma de todos los SalesAmount, el total de descuentos, a través de la suma de todos los DiscountAmount, y el total de cantidad de órdenes a través de la suma de todos los OrderQuantity. 
2. En el siguiente tab de Refresh Settings dejaremos todo por defecto.
3. En el siguiente tab, de Advanced settings, indicaremos la información de los aggregation groups del cubo, y pondremos los campos que se incluyen, si hay alguna jerarquía entre algunas de estos campos, y cuáles campos usualmente aparecen juntos, o tienen una relación uno del otro. Para ello, utilizaremos la siguiente configuración:
4. En los dos siguientes tabs dejamos toda la configuración por defecto y finalmente le damos a **save**.
5. Una vez creado el cubo, podemos ir a la sección principal de modelos, y veremos el cubo, y en la sección de **actions** damos click, y le damos click a Build para construir el cubo.
6. Una vez hecho esto, podemos ir a la sección de **Monitor** donde podemos monitorear el avance de la construcción del cubo como se observa a continuación:

## **Reportes:**

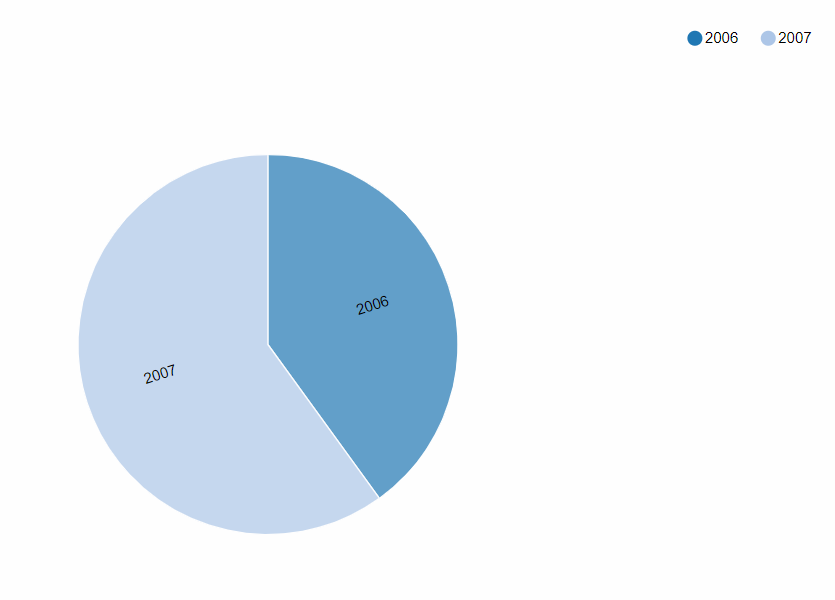
### **Reporte con Roll up:**

Se desea conocer el total de ventas mensuales de los años 2006 y 2007.

#### Sentencia SQL



#### Visualización de los datos

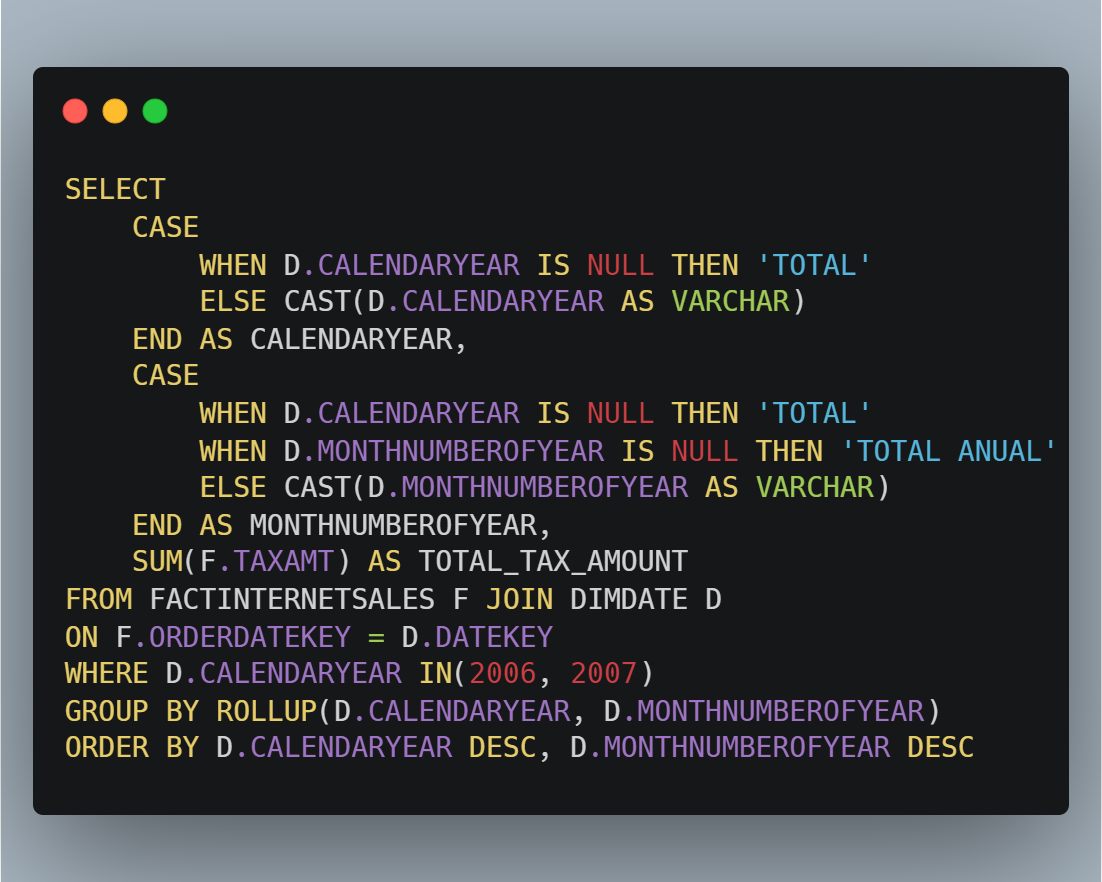


### 

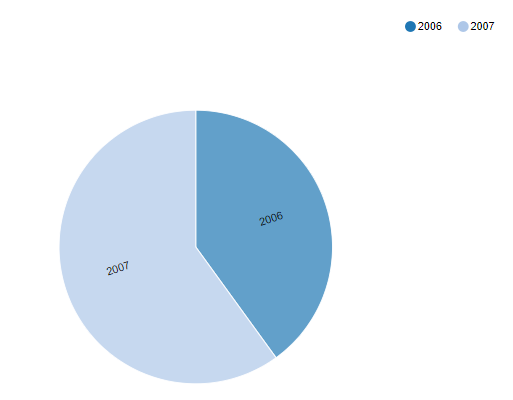
### **Reporte con Drill down:**

Se desea conocer el total de impuestos cobrados mensuales de los años 2006 y 2007.

#### Sentencia SQL



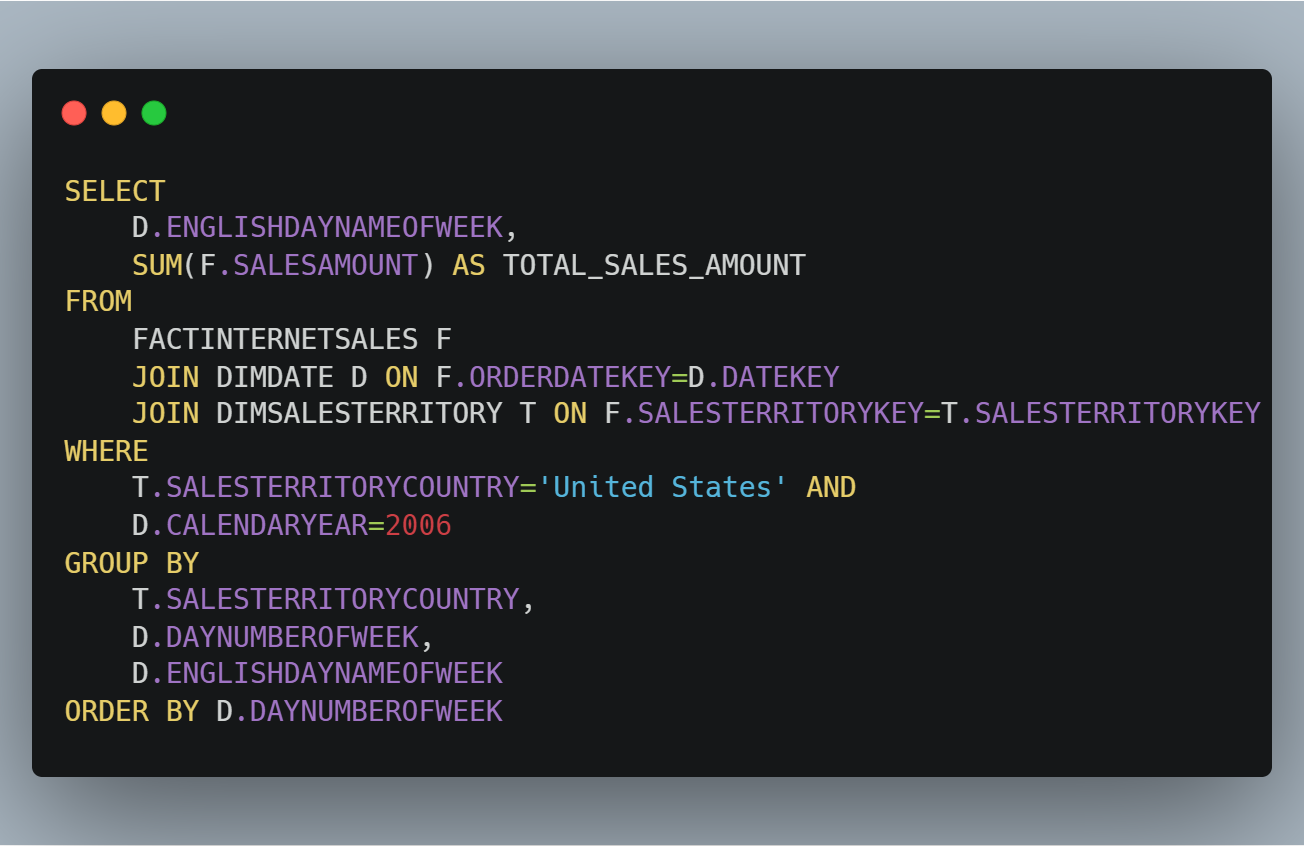
#### Visualización de los datos



### **Reporte con Dice:**

Se desea conocer el total de ventas por días que hubo en Estados Unidos en el año 2006.

#### Sentencia SQL



#### Visualización de los datos

