Texto

Descripción generada automáticamente

**ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL**

**Integrantes:**

García Quimi Nicole Marisol

Gutiérrez Plaza Camilo Andrés

**Tema:**

Cubo de información: Adventure works

**Profesora:**

PhD. Vanessa Echeverría

**Término:**

2021-2

1. **Objetivos:**

* Importar datos de una base de datos relacional (Mysql) al sistema de archivos Hadoop (HDFS) utilizando Sqoop[1].
* Crear una base de datos en el sistema de archivos Hadoop utilizando Hive[2].
* Crear un modelo de datos multidimensional para representar un proceso de negocio utilizando Apache Kylin[3].
* Crear un cubo de datos para la creación de reportes del proceso de negocio utilizando Apache Kylin.

1. **Descripción del dataset**

El dataset contiene información de la empresa Adventureworks que se dedica a vender bicicletas y partes de bicicletas a clientes y distribuidores. Dentro del dataset se encuentra la información de productos, ventas, y ubicación geográfica de las ventas.

1. **Instalación de Apache Kylin (ver 3.1.3)**

**Requisitos:**

* JDK 1.8
* Hadoop 2.7.0
* Hive 1.2.1
* Hbase 1.1.2 (with Zookeeper)
* Spark 2.3.1
* Kafka 1.1.1
* MySQL 5.1.73

Para este proyecto se instalará Apache Kylin dentro de un contenedor con los requisitos preinstalados por lo que es necesario la instalación de Docker[4].

Para descargar la imagen desde Docker, utilizando la consola escribimos:

docker pull apachekylin/apache-kylin-standalone:3.1.0

Luego de que se haya descargado la imagen correctamente corremos el contendor utilizando los siguientes parámetros:

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

El container\_id se imprime después de la ejecución del comando.

Cuando se haya inicializado ejecutamos la siguiente línea para entrar a la consola del contenedor:

docker exec -it <container\_id> bash

1. **Importación de datos**

Para importar los datos desde nuestro dataset que se encuentra en una base de datos relacional tenemos que instalar Sqoop para cargar el esquema relacional al HDFS.

**Importar datos a Mysql**

Para importar los datos del dataset a nuestra base de datos utilizaremos el script sql que haya proporcionado nuestro instructor.

Utilizando el servidor de mysql corriendo accedemos a la consola:

mysql -u root -p

la contraseña del usuario root es: 123456

Luego de entrar ejecutamos el script sql desde la consola:

source <script\_path>

**Instalación de Apache Sqoop**

Para instalar Apache Sqoop debemos descargar los archivos binarios desde el repositorio de Apache:

wget http://archive.apache.org/dist/sqoop/1.4.7/sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz

Una vez descargados descomprimimos la carpeta de Apache Sqoop y la movemos a las librerías del usuario:

tar -xvf sqoop-1.4.7.bin\_\_hadoop-2.6.0.tar.gz

mv sqoop-1.4.7.bin\_\_hadoop-2.6.0 /usr/lib/sqoop

Configuramos el Bashrc de nuestro ambiente para poder ejecutar sqoop desde la línea de comandos:

export SQOOP\_HOME=/usr/lib/sqoop

export PATH=$PATH:$SQOOP\_HOME/bin

source ~/.bashrc

Configuramos Sqoop para que apunte al Hadoop instalado en nuestro contenedor:

cd $SQOOP\_HOME/conf

mv sqoop-env-template.sh sqoop-env.sh

Abrimos el archivo sqoop-env.sh y editamos las siguientes líneas:

export HADOOP\_COMMON\_HOME=/home/admin/hadoop-2.7.0

export HADOOP\_MAPRED\_HOME=/home/admin/hadoop-2.7.0

**Instalación de Mysql Connector**

Una vez configurado, descargamos el conector de mysql para que Sqoop se conecte a nuestra base de datos relacional:

wget http://ftp.ntu.edu.tw/MySQL/Downloads/Connector-J/mysql-connector-java-5.1.49.tar.gz

Una vez descargado lo descomprimimos y añadimos a las dependencias de Sqoop.

tar -zxf mysql-connector-java-5.1.49.tar.gz

cd mysql-connector-java-5.1.49

mv mysql-connector-java-5.1.49-bin.jar /usr/lib/sqoop/lib

**Carga de datos a HDFS**

Para cargar los datos al sistema de archivos Hadoop debemos conectarlo con nuestra base de datos relacional subiendo el esquema como archivos parquet a través de Sqoop utilizando la siguiente línea:

sqoop import-all-tables --connect jdbc:mysql://localhost:3306/adventureworksdw --username root -P -m 1

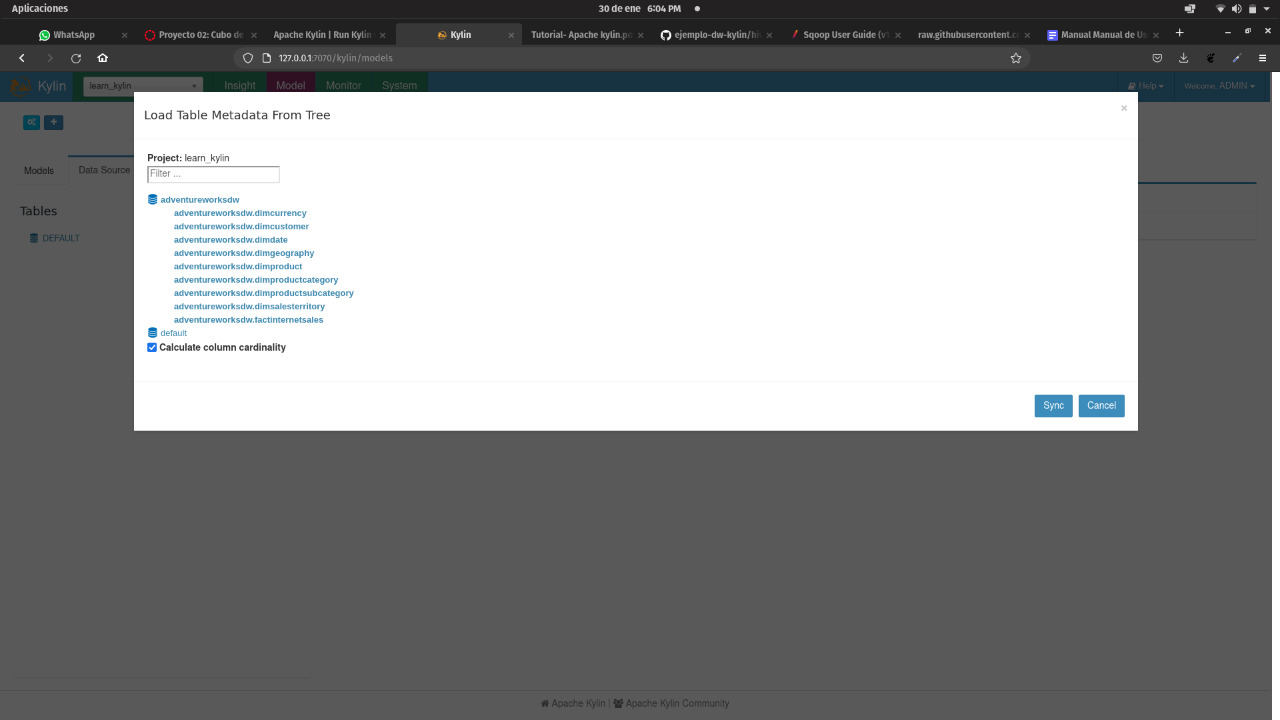
**Crear la base de datos con Hive**

Para crear la base de datos de hive se debe ejecutar el script hql proporcionado por su instructor. Deben de cambiar el location que se encuentra en el script para que concuerde con su directorio en el HDFS:

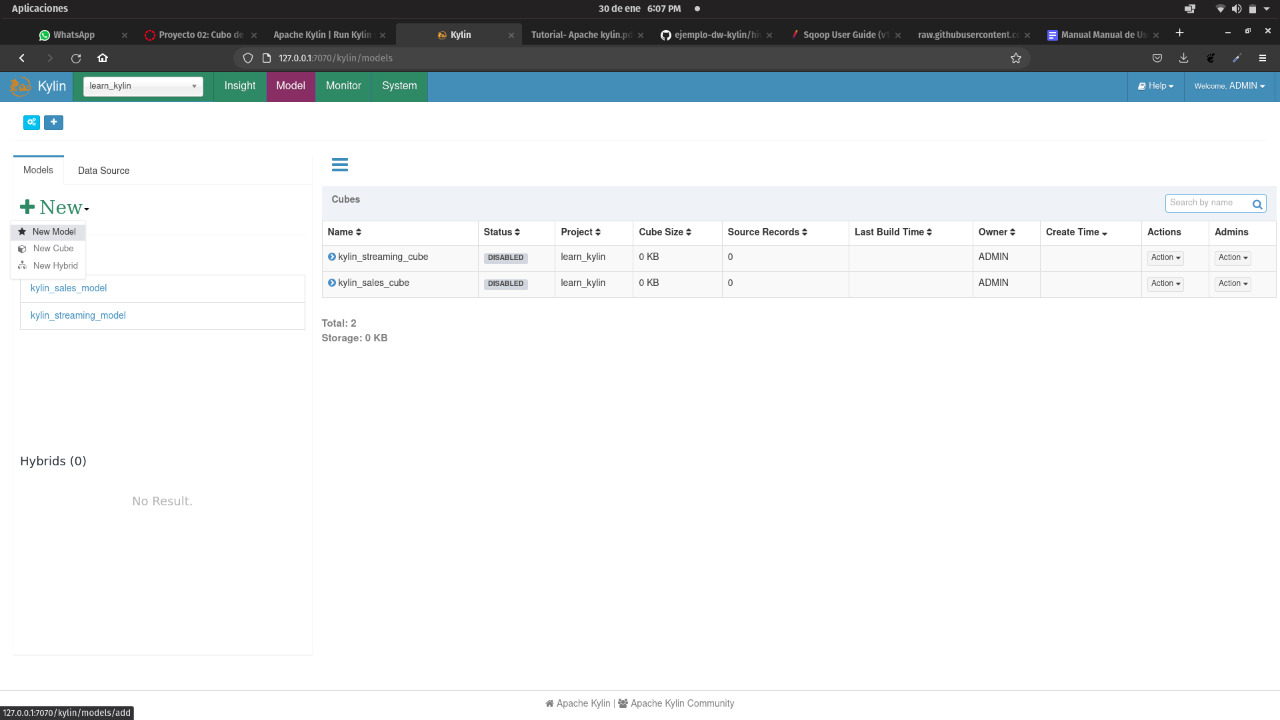
hive -f <script\_path>

1. **Creación del modelo**

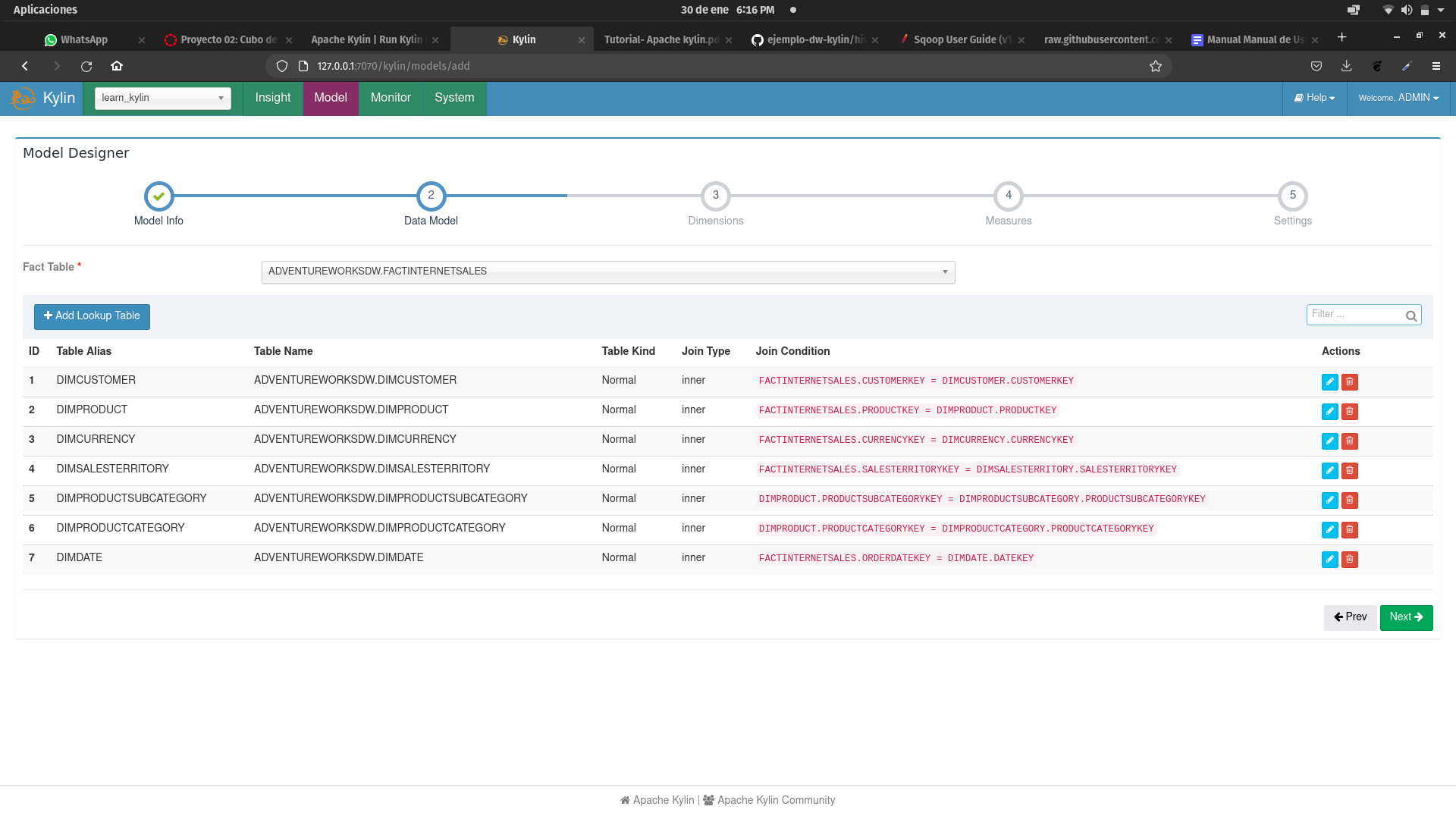
Para la creación del modelo es necesario la sincronización de la base de datos de Hive en Kylin:



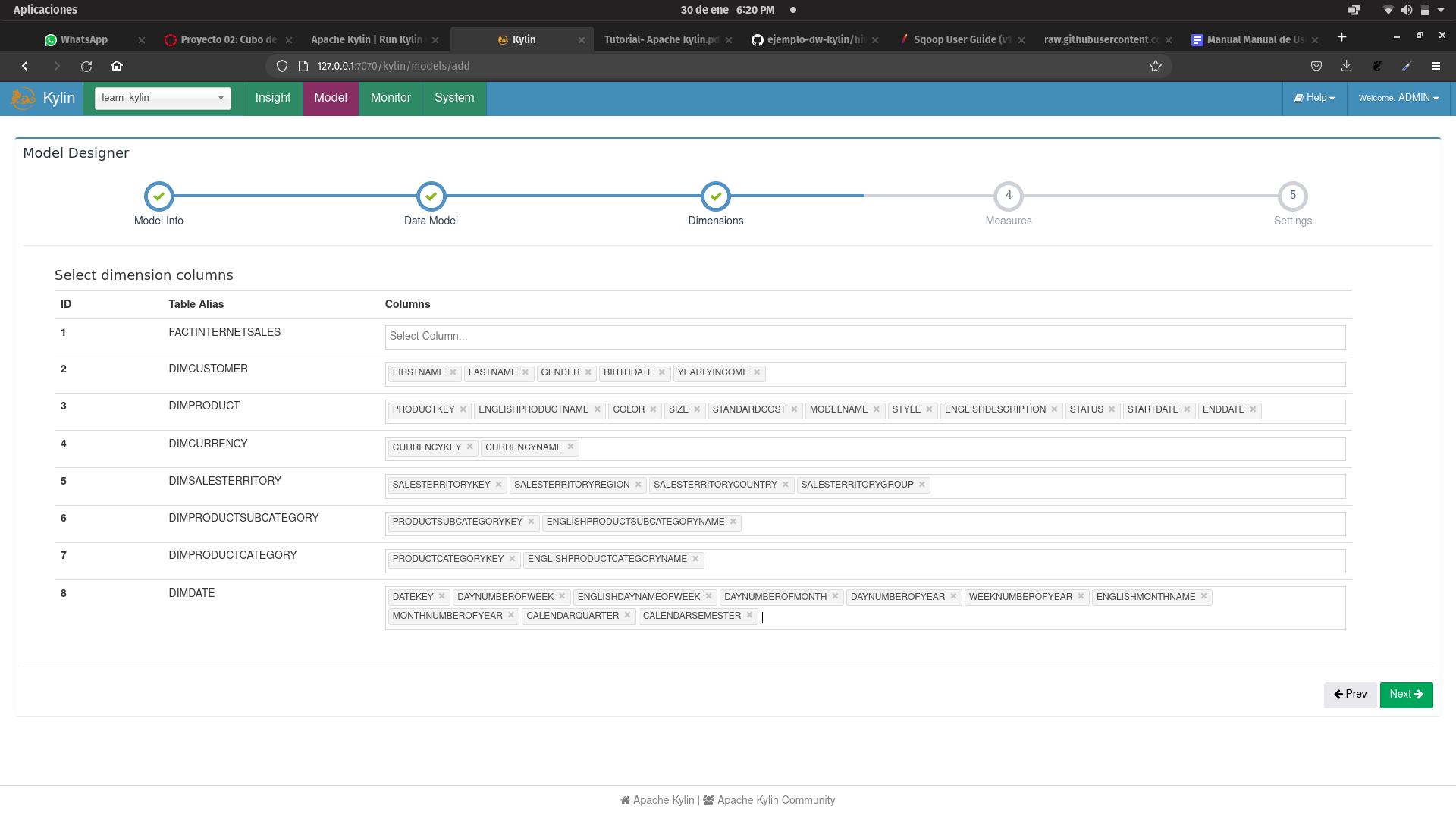
Luego de sincronizar los datos en Apache Kylin procedemos a la creación del modelo en la sección **Model**.



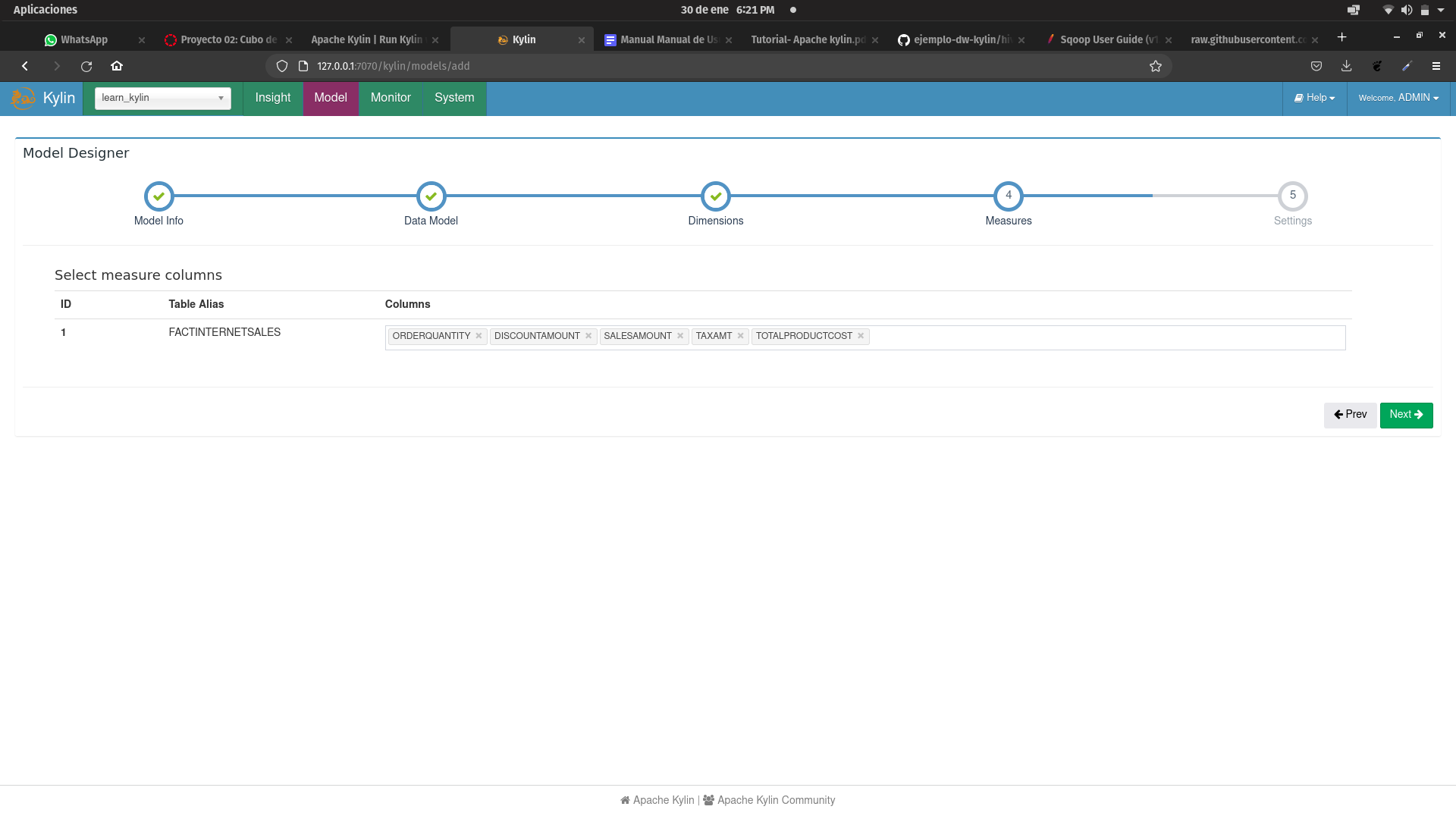
En la sección **data model** añadimos los Join entre la tabla de hechos y las tablas de dimensiones.



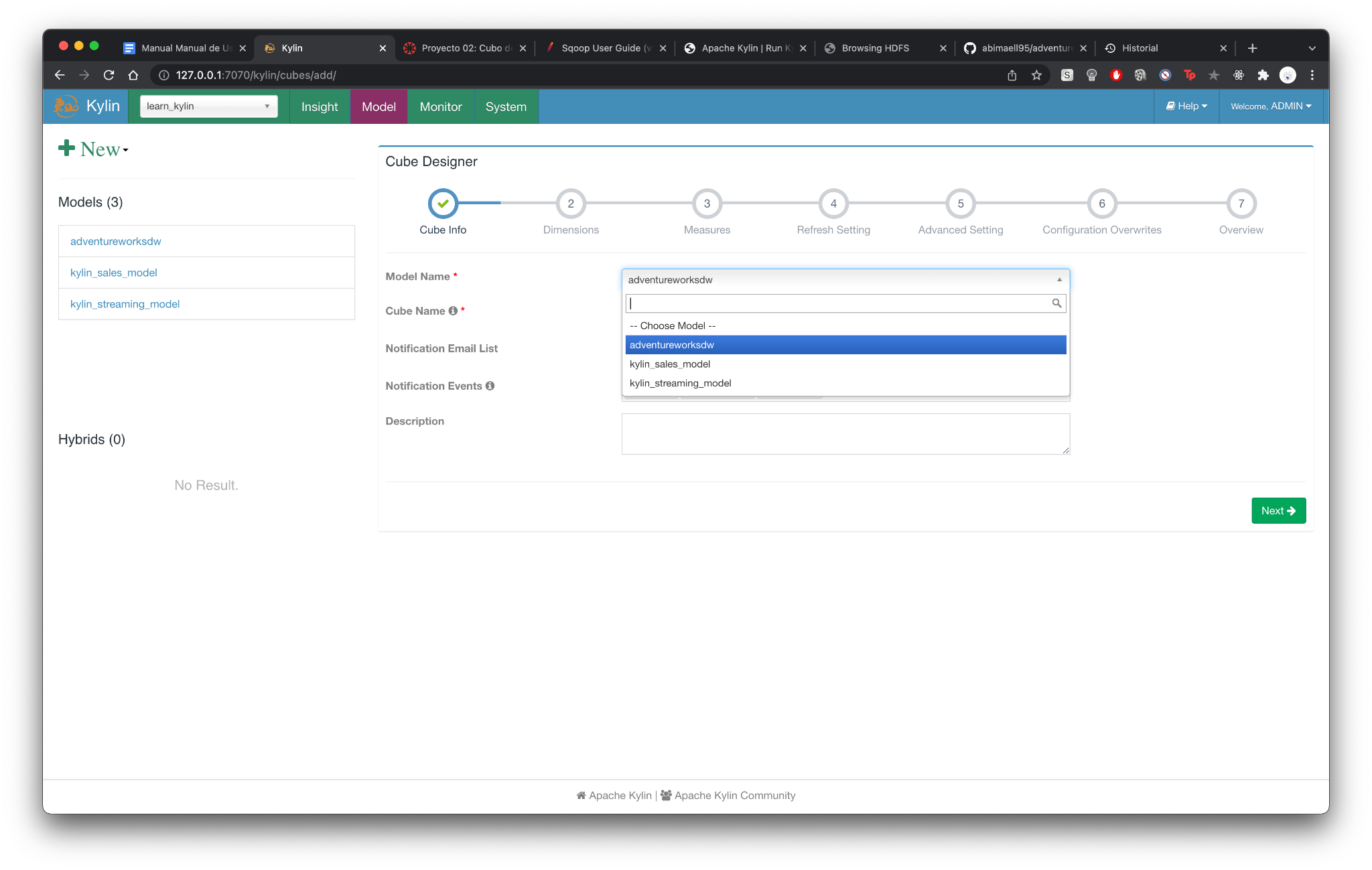
En la sección **dimensions** añadimos los campos que van en cada una de las dimensiones.



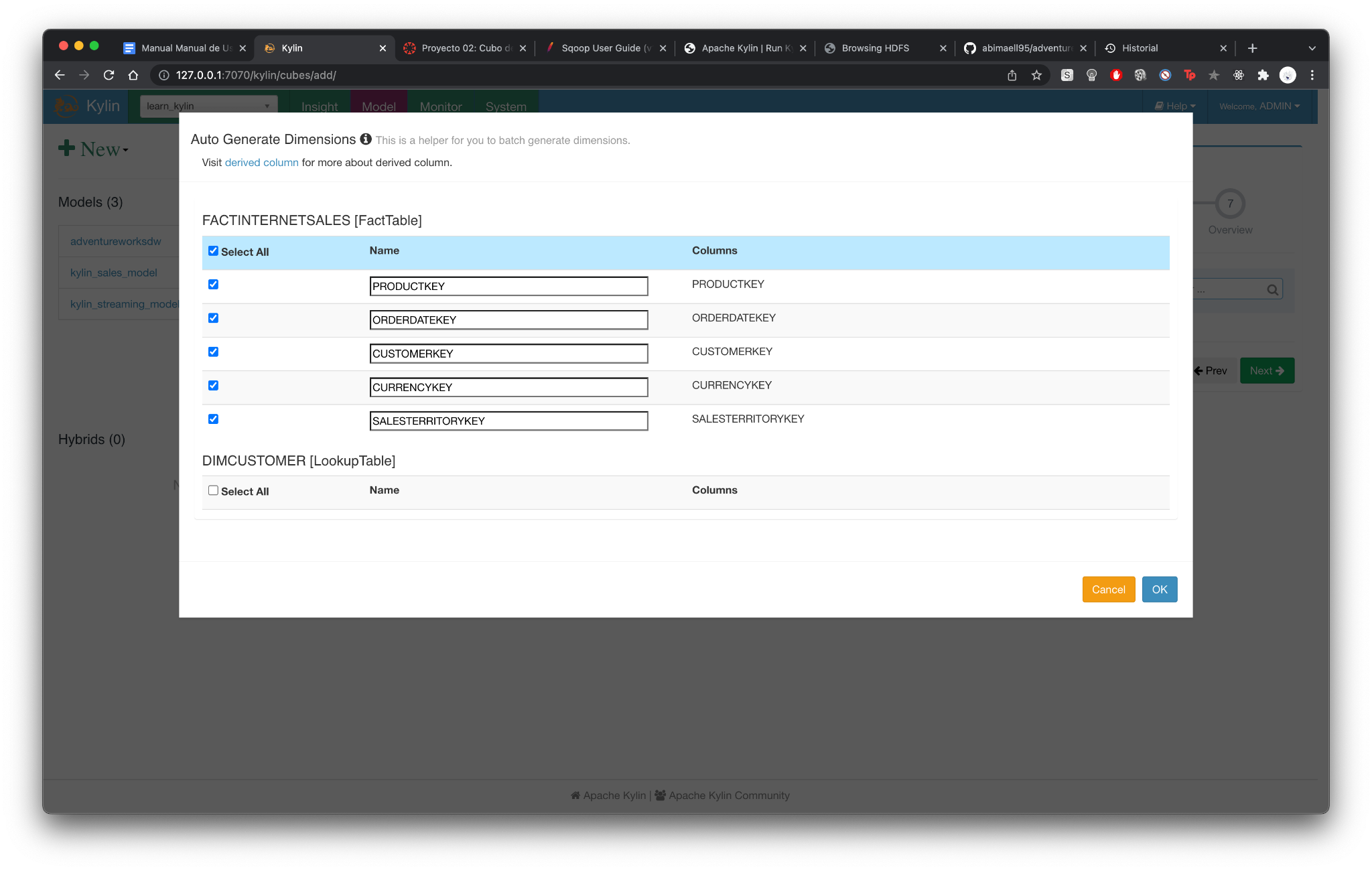
Finalmente en la sección **measures** añadimos los campos seleccionados de la tabla de hechos.



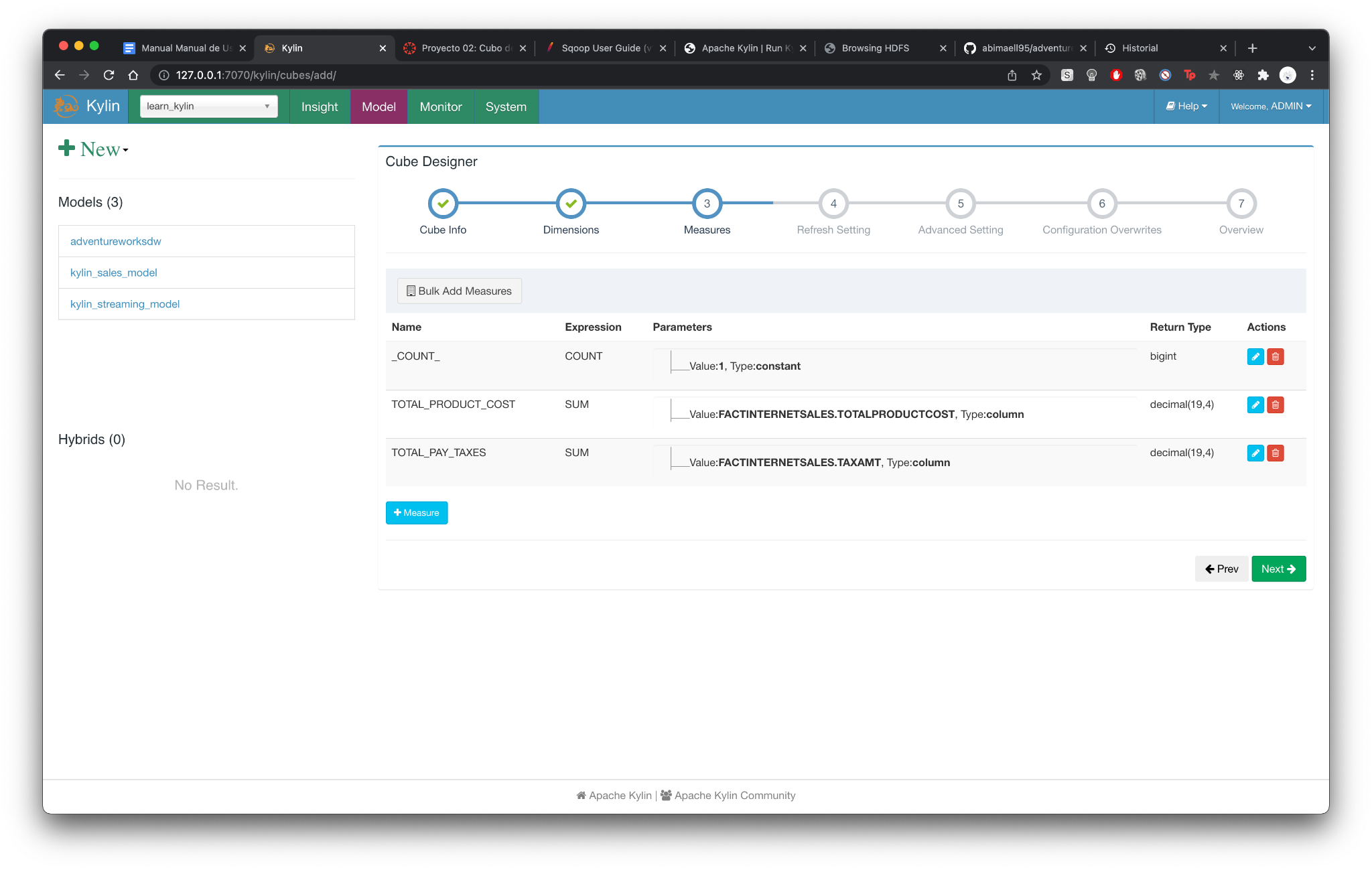
1. **Creación del Cubo**

Para la creación del cubo es necesario que nuestro modelo se haya creado satisfactoriamente.

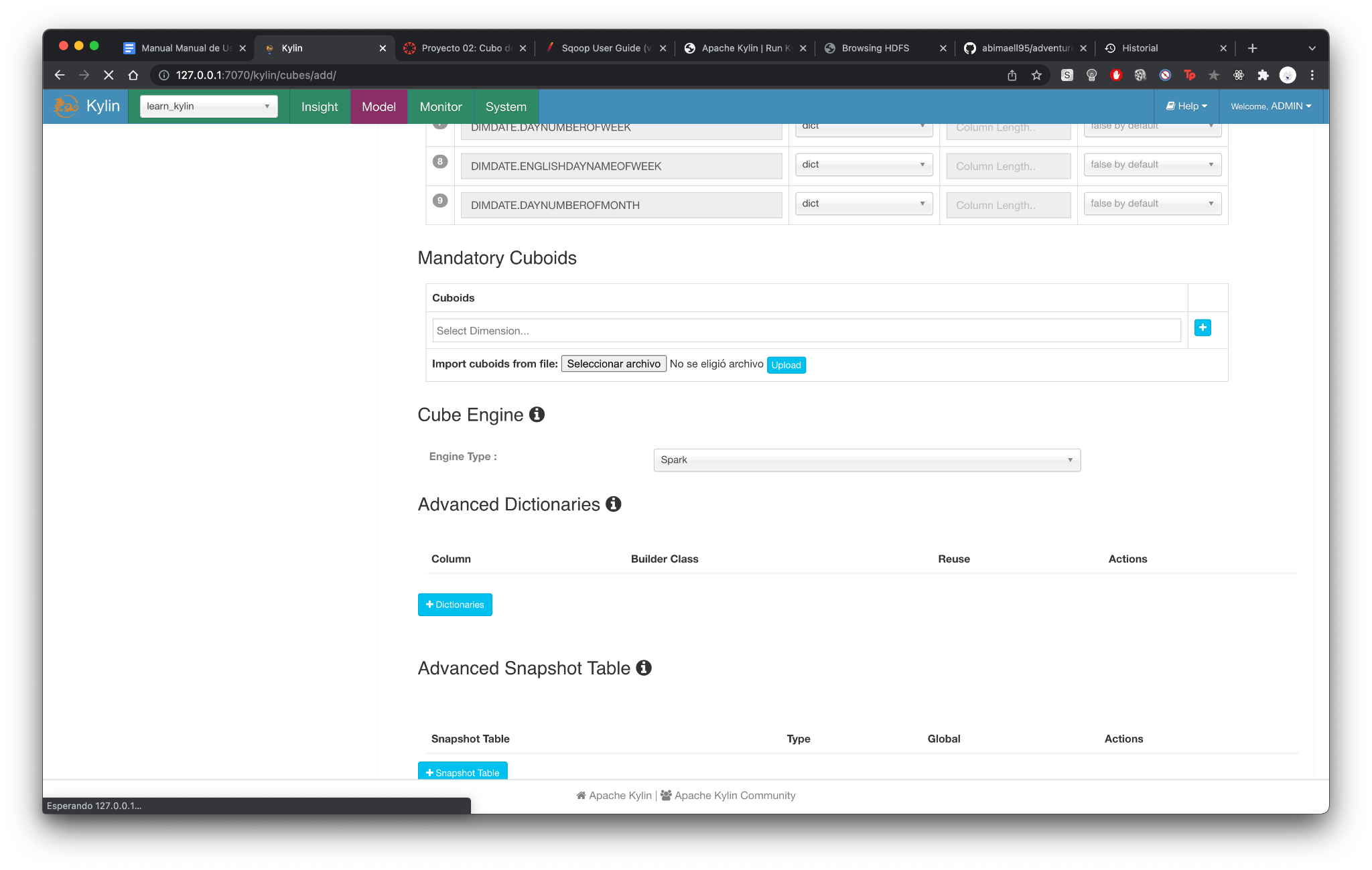
En la sección de **dimensions** seleccionamos los campos que serán utilizados en las dimensiones seleccionadas.



En la sección **measures** añadimos las funciones de agregación para nuestro proceso de negocio.

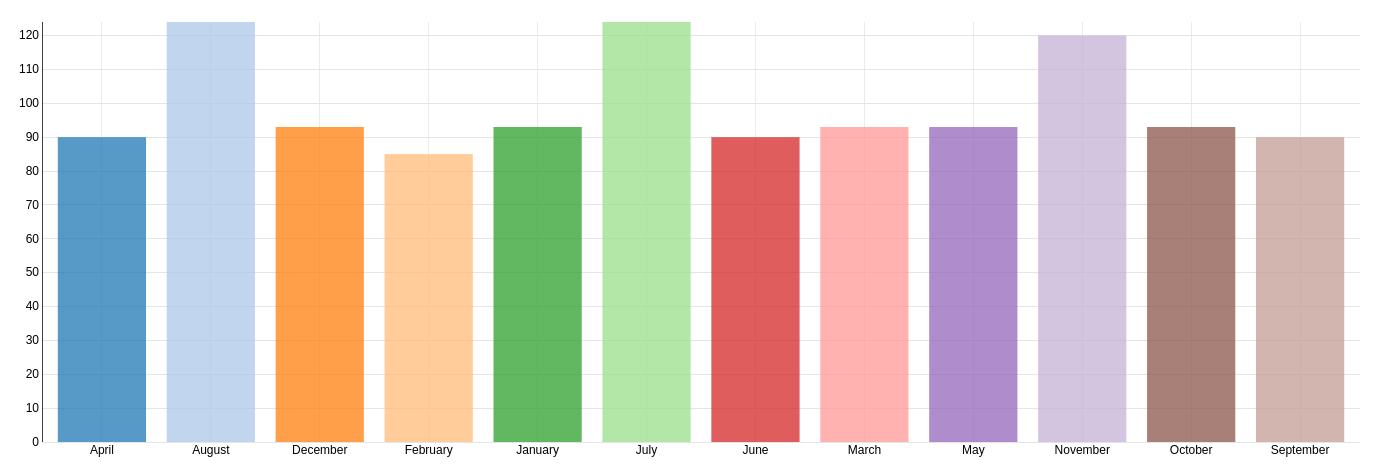


Finalmente en la sección **Advanced Setting** cambiamos el motor del cubo por **Spark**.

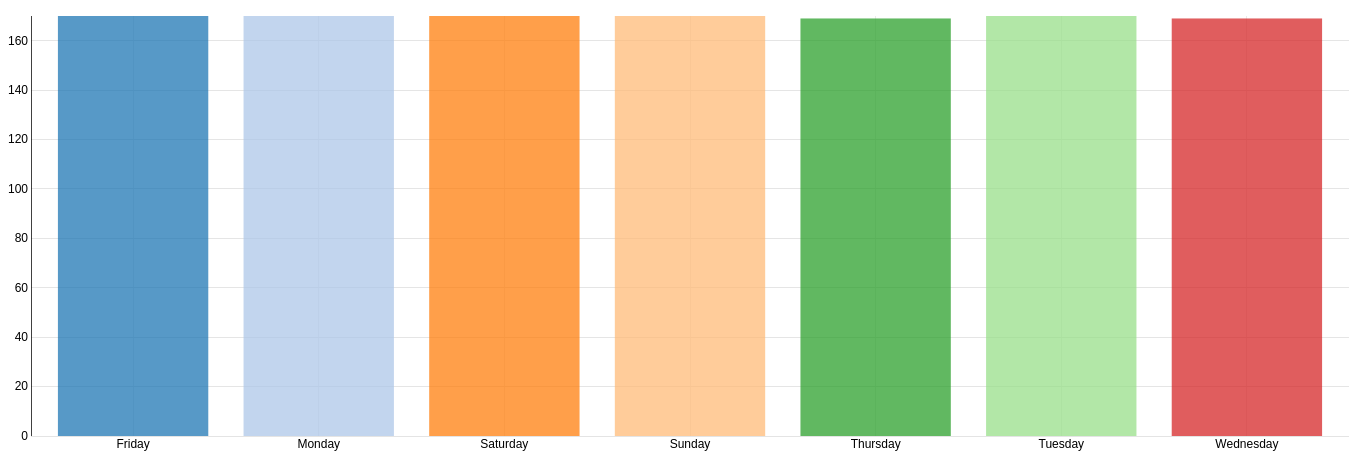


1. **Reportes**

Total de ventas por mes

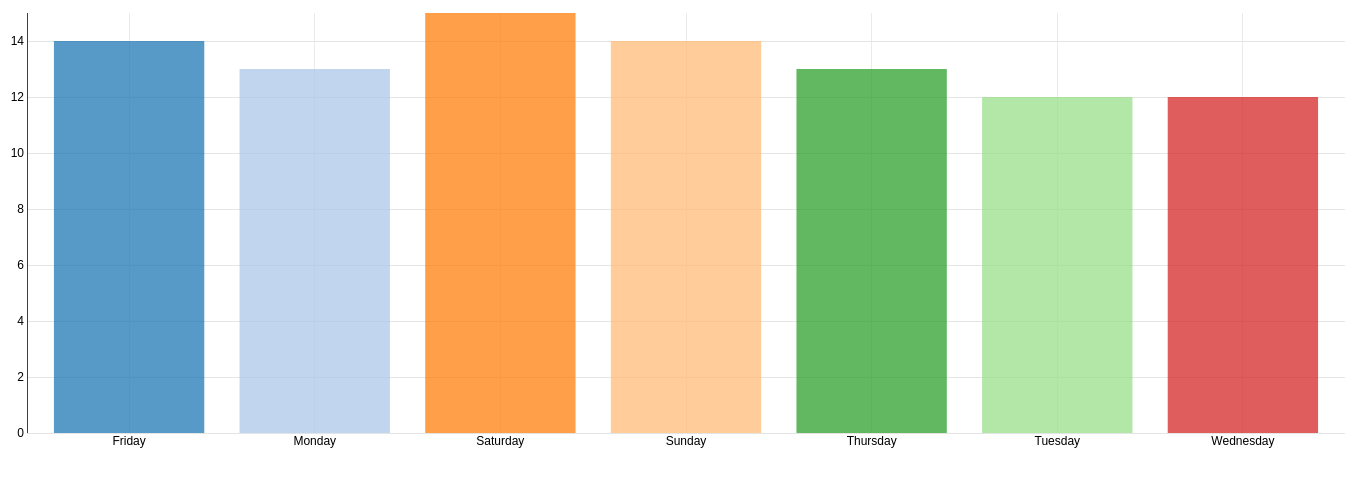
select englishmonthname, count(englishmonthname) total from DIMDATE group by englishmonthname

Total de ventas por día de la semana

select englishdaynameofweek, count(englishdaynameofweek) as total from dimdate d group by englishdaynameofweek

Total de ventas por día de la semana en el mes de diciembre

select englishdaynameofweek, count(englishdaynameofweek) as total from dimdate where monthnumberofyear=12 group by englishdaynameofweek



1. **Referencias**

[1] Apache Sqoop https://sqoop.apache.org/docs/1.4.7/SqoopUserGuide.html

[2] Apache Hive https://cwiki.apache.org/confluence/display/HIVE

[3] Apache Kylin

https://kylin.apache.org/docs/tutorial/web.html

[4] Docker

https://docs.docker.com/