Carga de la base de datos con PostgreSQL

Instalación

Para realizar la instalación de este sistema de gestión de bases de datos relacional orientado a objetos y libre, primero hemos procedido a descargar el paquete de instalación a través de la página de la compañía **EnterpriseDB**, que proporciona software y servicios basados en la base de datos de código abierto PostgreSQL: https://www.enterprisedb.com/downloads/postgres-postgresql-downloads. En nuestro caso, hemos procedido a instalar la versión 9.6.6 para el sistema operativo Linux x86-64.

Conexión a la base de datos

Para conectarnos a la base de datos PostgreSQL, hemos procedido a utilizar el programa terminal interactivo proporcionado por PostgreSQL denominado **psql**.

Base de datos con los datos originales

Antes que nada, hemos creado una base de datos con los datos originales que nos proporciona la competición KDD Cup 2017. Para poder construirla, primero nos hemos conectado a la base de datos de esta forma:

```
$ psql tfgdatosoriginales
```

Para cargar los datos originales de la competición en la base de datos PostgreSQL, hemos desarrollado un script que realiza la carga de todas estas tablas de forma directa. Para ejecutarlo, en el terminal se escribe lo siguiente:

```
$ \i [ruta_del_script]
```

Tabla *road_links ("links_table3.csv")*

El esquema de la tabla original proporcionado por la competición es la siguiente:

Campo	Tipo	Descripción
link_id	string (char(3))	Identificador del enlace
length	float	Longitud del enlace en metros
width	float	Anchura del enlace en metros

lanes	int	Número de carriles
in_top	string (varchar(7))	Este atributo contiene los enlaces entrantes al enlace actual, separados por comas
out_top	string (varchar(7))	Este atributo contiene los enlaces salientes del enlace actual, separados por comas
lane_width	float	Anchura de cada uno de los carriles del enlace en metros

Esta tabla contiene la descripción de cada uno de los enlaces que puede formar una carretera.

Tabla *vehicle_routes ("routes_table4.csv")*

El esquema de la tabla original proporcionado por la competición es la siguiente:

Campo	Tipo	Descripción
intersection_id	string (char(1))	Identificador de la intersección
tollgate_id	string (char(1))	Identificador de la barrera de peaje
link_seq	string (varchar(47))	Secuencia de enlaces que conforman la ruta desde la intersección hasta la barrera de peaje

La red de carreteras utilizada en la competición es un grafo dirigido formado por enlaces de carreteras interconectados. Una ruta en la red está representada por una secuencia de enlaces. Para cada enlace de la carretera, el tráfico de vehículos proviene de uno o más "enlaces viales entrantes" y entra en uno o más "enlaces viales salientes".

Tabla *vehicle_trajectories_training ("trajectories_table 5_training.csv")*

El esquema de la tabla original proporcionado por la competición es la siguiente:

Campo	Tipo	Descripción
intersection_id	string (char(1))	Identificador de la intersección
tollgate_id	string (char(1))	Identificador de la barrera de peaje
vehicle_id	string (varchar(30))	Identificador del vehículo
starting_time	datetime (timestamp)	Momento del tiempo en el que el vehículo entra en la ruta
travel_seq	string (varchar(400))	Trayectoria de la ruta formada por un conjunto de enlaces. Estos enlaces están separados por un ";" y, para cada enlace, se especifica, separados por "#", su identificador, el momento del tiempo en el que el vehículo entra en ese enlace y el tiempo que pasa el vehículo atravesando dicho enlace en segundos.
travel_time	float	Tiempo total que tarda el vehículo en viajar desde la intersección hasta la barrera de peaje.

Esta tabla contiene cada uno de los vehículos que ha viajado en algún momento, entre el 19 de Julio y el 17 de Octubre, por alguna de las rutas establecidas en la tabla *vehicle_routes*. Para cada vehículo se establece el momento en el que entró en una ruta, el tiempo que estuvo ese vehículo en cada uno de los enlaces que forma dicha ruta y el tiempo total que tarda en realizar esa ruta.

Tabla *traffic_volume_tollgates_training (volume_table6_training.csv")*

El esquema de la tabla original proporcionado por la competición es la siguiente:

Campo	Tipo	Descripción
time	datetime (timestamp)	Momento en el que un vehículo atraviesa la barrera de peaje
tollgate_id	string (char(1))	Identificador de la barrera de peaje
direction	string (char(1))	Dirección en la que el vehículo atraviesa la barrera de peaje
vehicle_model	int	Modelo del vehículo. Este número (compendido entre los valores 0 y 7), cuanto mayor sea, mayor es su capacidad
has_etc	string (char(1))	Indica si el vehículo utiliza un ETC (Electronic Toll Collection)
vehicle_type	string(char(1))	Tipo de vehículo. Indica si el vehículo es de pasajeros o de carga

En esta tabla se registran todos los vehículos que han pasado por alguna barrera de peaje situada en la topología de carreteras proporcionada por la competición. Con respecto al atributo *vehicle_type*, la competición no proporciona esta columna, por lo que no se considera importante.

Tabla *traffic_volume_tollgates_training (volume_table6_training.csv")*

El esquema de la tabla original proporcionado por la competición es la siguiente:

Campo	Tipo	Descripción
date	date	Fecha
hour	int	Hora

pressure	float	Presión del aire (*hPa*)
sea_pressure	float	Presión del nivel del mar (*hPa*)
wind_direction	float	Dirección del viento (*º*)
wind_speed	float	Velocidad del viento (*m/s*)
temperature	float	Temperatura(*ºC*)
rel_humidity	float	Humedad relativa
precipitation	float	Precipitaciones (*mm*)

Esta tabla contiene los datos meteorológicos de cada una de las fechas contenidas en intervalos de 3 horas dentro del conjunto de entrenamiento.

COMPROBACIONES DE LOS DATOS

- Comprobar si las trayectorias formadas por un conjunto de enlaces en la tabla *vehicle_trajectories_training* coincide con alguna de las trayectorias definidas en la tabla *vehicle_routes*.
- Comprobar si el tiempo total de cada uno de los vehículos corresponde a la suma de todos los tiempos de cada uno de los enlaces de la ruta por la que pasa el vehículo en la tabla *vehicle_trajectories_training*.

ERRORES

- En la tabla *traffic_volume_tollgates_training*, el campo **vehicle_type** no tienen ningún valor en ninguna fila.