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:

En nuestro caso, hemos procedido a instalar la versión 9.3.20 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 poder construirla, nos tenemos que conectar a la base de datos utilizando el terminal interactivo que nos proporciona PostgreSQL:

$ \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 forman 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. Si es 0, la dirección es de salida; si es 1, la dirección es de entrada.
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 *weather_data ("weather (table 7)_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.

Nota: En meteorología, es importante tener en cuenta que la dirección nos indica de dónde viene el viento, no hacia dónde va. Se mide en grados, desde 0º (excluido) hasta 360º (incluido), girando en el sentido de las agujas del reloj en el plano horizontal visto desde arriba. Valores cercanos a 1º y 360º indican viento del norte, cercanos a 90º viento del este, 180º del sur y 270º del oeste. Entre estos valores tendremos el resto de direcciones: nordeste, sureste, suroeste y noroeste.

Tabla *travel_time_intersection_to_tollgate ("trajectories_table5_training_20min_avg_travel_time.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
time_window	string(varchar(43))	Ventana de tiempo de 20 minutos
avg_travel_time	float	Tiempo medio de viaje (segundos)

Esta tabla se obtiene como resultado de la ejecución del script ***aggregate_travel_time.py*** sobre la tabla ***vehicle_trajectories_training***. Para construir esta tabla, el script crea un diccionario para guardar el tiempo de viaje para cada una de las rutas por cada ventana de tiempo y calcula la media del tiempo de viaje para cada una de esas rutas por cada ventana de tiempo. Esta tabla es el formato de tabla requerido para visualizar las predicciones que se realicen del tiempo medio de viaje por cada una de las rutas y ventana de tiempo de 20 minutos. En este caso, la tabla cargada en la base de datos es un ejemplo de creación de ese tipo de tablas.

Tabla *traffic_volume_tollgates ("volume_table 6_training_20min_avg_volume.csv")*

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

Campo	Tipo	Descripción
tollgate_id	string (char(1))	Identificador de la intersección
time_window	string (varchar(45))	Ventana de tiempo de 20 minutos
direction	string(char(1))	Dirección en la que se atraviesa la barrera de peaje
volume	int	Número de vehículos que atraviesan la barrera de peaje en la ventana de tiempo de 20 minutos

Esta tabla se obtiene como resultado de la ejecución del script ***aggregate_volume.py*** sobre la tabla ***volume_table 6_training.csv***. Para construir esta tabla, el script crea un diccionario para guardar el volumen de vehículos que atraviesa cada una de las barreras de peaje cada ventana de tiempo de 20 minutos. Esta tabla es el formato de tabla requerido para visualizar las predicciones que se realicen del volumen de vehículos que atraviesa cada una de las barreras de peaje en cada ventana de tiempo de 20 minutos. En este caso, la tabla cargada en la base de datos es un ejemplo de creación de ese tipo de tablas.

Base de datos con los datos modificados

Para manejar los datos proporcionados de una forma más manejable y coherente, hemos procedido a realizar las modificaciones oportunas de los tipos de datos de las columnas de las distintas tablas. Para poder construirla, nos tenemos que conectar a la base de datos utilizando el terminal interactivo que nos proporciona PostgreSQL:

$ psql tfgdatosmodificados

Para realizar la carga de la base de datos modificados en 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]

Todas las tablas modificadas se han cargado a partir de los datasets originales, con el objetivo de conservar intactos los datos proporcionados.

Tabla *road_links_modified ("links_table3.csv")*

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
link_id	smallint	Identificador del enlace
length	float	Longitud del enlace en metros
width	float	Anchura del enlace en metros
lanes	int	Número de carriles
in_top	smallint[]	Este atributo contiene los

		enlaces entrantes al enlace actual, separados por comas
out_top	smallint[]	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

Uno de los cambios que se ha realizado ha sido cambiar el tipo de dato de la columna *link_id* a **smallint**. Por otra parte, los tipos de las columnas *in_top* y *out_top* se han establecido como **arrays de smallint** puesto que sus valores son conjuntos de enlaces. Para poder cargar desde el archivo .csv la tabla y que no surgiera conflicto de tipos, inicialmente se han creado las columnas con el tipo **varchar**, se han creado arrays a partir de los valores de esas columnas y se ha alterado la tabla para realizar una conversión explícita a arrays de tipo **smallint**.

Tabla *vehicle_routes_modified ("routes_table4.csv")*

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
intersection_id	char(1)	Identificador de la intersección
tollgate_id	smallint	Identificador de la barrera de peaje
link_seq	smallint[]	Secuencia de enlaces que conforman la ruta desde la intersección hasta la barrera de peaje

En esta table se ha cambiado el tipo de la columna *tollgate_id* a **smallint** y la columna *link_seq* a **smallint[]**. Este último atributo se ha modificado de la misma forma que las columnas *in_top* y *out_top* de la tabla anterior.

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

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
intersection_id	char(1)	Identificador de la intersección
tollgate_id	smallint	Identificador de la barrera de peaje
vehicle_id	int	Identificador del vehículo
starting_time	timestamp	Momento del tiempo en el que el vehículo entra en la ruta
travel_seq	link_object[]	Trayectoria de la ruta formada por un conjunto de enlaces. Estos enlaces están representados con un objeto *link_object*, que contiene 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.

En esta tabla se han modificado los tipos de las columnas *tollgate_id* (a **smallint**), *vehicle_id* (a **int**) y *travel_seq* (a **link_object[]**). Esta última columna tiene un tipo compuesto, formado por los siguientes atributos:

- *id* : Identificador del enlace (**smallint**)
- *entrance_time*: Momento del tiempo en el que el vehículo entra en ese enlace (**timestamp**)
- *duration*: Tiempo que pasa el vehículo atravesando dicho enlace en segundos (**float**)

Para convertir los valores de la columna *travel_seq* a un conjunto de objetos de tipo **link_object[]**, primero se han convertido en arrays de **varchar** y se ha convertido el tipo de la columna a este tipo de arrays. A continuación, se ha creado un bloque con el objetivo de recorrer cada una de las filas de la tabla, de tal forma que, por cada fila, se coge el array de **varchar**, se crea un objeto **link_object** por cada uno de los elementos del array y se guardan estos objetos en un **link_object[]**. Así, se actualiza la tabla con este nuevo tipo y, para que la columna tenga el tipo correcto, se realiza la conversión explícita a **link_object[]**.

Tabla *traffic_volume_tollgates_training_modified ("volume_table6_training.csv")*

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
time	timestamp	Momento en el que un vehículo atraviesa la barrera de peaje
tollgate_id	smallint	Identificador de la barrera de peaje
direction	smallint	Dirección en la que el vehículo atraviesa la barrera de peaje. Si es 0, la dirección es de salida; si es 1, la dirección es de entrada.
has_etc	boolean	Indica si el vehículo utiliza un ETC (Electronic Toll Collection)

En esta tabla se han eliminado las columnas *vehicle_model* y *vehicle_type* puesto que no son relevantes a la hora de realizar las predicciones pertinentes. Por otra parte, se han modificado los tipos de los atributos *tollgate_id* (a **smallint**), *direction* (a **smallint**) y *has_etc* (a **boolean**).

Tabla *weather_data_modified ("weather (table 7)_training.csv")*

El esquema de la tabla no se ha modificado debido a que los tipos de las columnas en la tabla original son los correctos. Sin embargo, se han modificado aquellas filas en las que la columna *wind_direction* tenía el valor **999017** puesto que el valor que admite este atributo son *grados* (º) y el valor debe estar entre 0 y 360 grados. La modificación que se ha procedido a realizar fue realizar la media entre la dirección del viento del día anterior y del dia posterior, con el objetivo de realizar una aproximación y no eliminar completamente una fila; nos conviene tener todos los días de los que dispongamos.

Tabla *travel_time_intersection_to_tollgate_modified ("trajectories_table5_training_20min_avg_travel_time.csv")*

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
intersection_id	char(1)	Identificador de la intersección
tollgate_id	smallint	Identificador de la barrera de peaje
time_window	timestamp ARRAY[2]	Ventana de tiempo de 20 minutos
avg_travel_time	float	Tiempo medio de viaje (segundos)

En esta tabla se han modificado los tipos de las columnas *tollgate_id* (a **smallint**) y *time_window* (a **timestamp ARRAY[2]**). El proceso llevado a cabo para convertir esta última columna al nuevo tipo es similar al que se ha llevado a cabo para el tipo **link_object[]** en la tabla *vehicle_trajectories_training_modified*. Para representar el intervalo en la columna *time_window* se ha utilizado un array de dos posiciones por conveniencia.

Tabla *traffic_volume_tollgates_modified ("volume_table 6_training_20min_avg_volume.csv")*

El nuevo esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
tollgate_id	smallint	Identificador de la intersección
time_window	timestamp ARRAY[2]	Ventana de tiempo de 20 minutos
direction	smallint	Dirección en la que se atraviesa la barrera de peaje
volume	int	Número de vehículos que atraviesan la barrera de peaje en la ventana de tiempo de 20 minutos

En esta tabla se han modificado los tipos de las variables *tollgate_id* (a **smallint**), *time_window* (a **timestamp ARRAY[2]**) y *direction* (a **smallint**). Para realizar el cambio del tipo del atributo *time_window* se ha realizado el mismo proceso que la columna *time_window* en la tabla anterior.

Base de datos con los datos relacionados con la primera fase de prueba

Para separar la primera fase de entrenamiento de la primera fase de pruebas y estructurar mejor la información, se ha procedido a crear una nueva base de datos para manejar los datos proporcionados por la competición para la primera fase de pruebas. Para poder construirla, nos tenemos que conectar a la base de datos utilizando el terminal interactivo que nos proporciona PostgreSQL:

$ psql tfgtest1

Para realizar la carga de la base de datos modificados en 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 *travel_time_intersection_to_tollgate_test1 ("test1_20min_avg_travel_time.csv")*

El esquema de esta tabla es la siguiente:

Campo	Tipo	Descripción
intersection_id	char(1)	Identificador de la intersección
tollgate_id	smallint	Identificador de la barrera de peaje
time_window	timestamp ARRAY[2]	Ventana de tiempo de 20 minutos
avg_travel_time	float	Tiempo medio de viaje (segundos)

Para comprender el sentido de esta relación, es necesario observar la siguiente imagen:

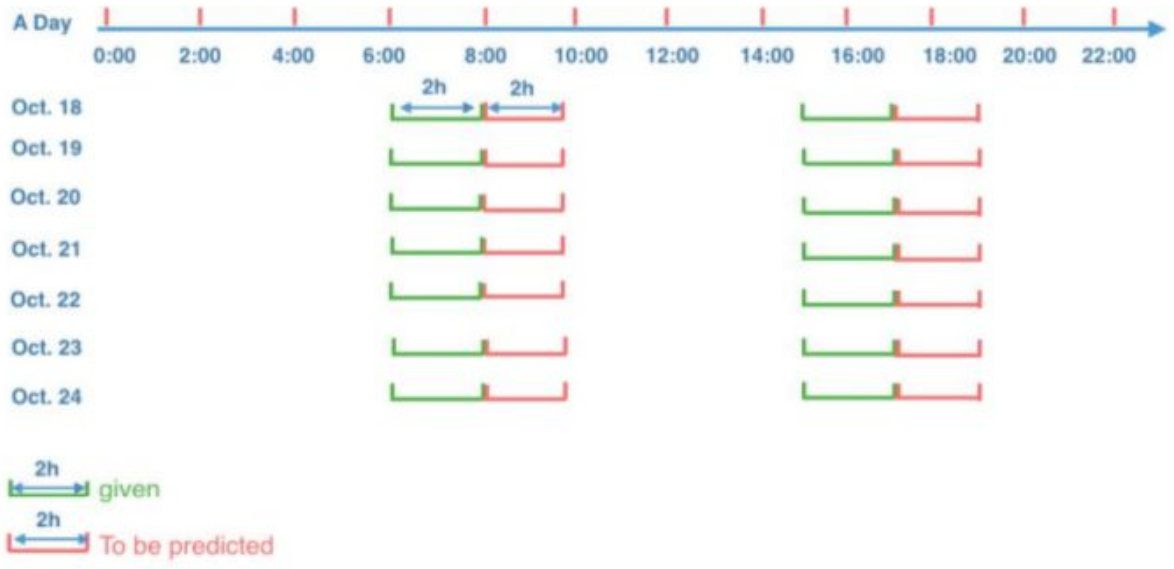

Figura 1. Ventanas de tiempo para la predicción de tráfico

En esta tabla se proporciona el tiempo promedio de viaje de cada una de las rutas en los intervalos de tiempo previos (colo verde) a los intervalos de tiempo a predecir. Particularmente, se especifica el tiempo promedio de viaje en intervalos de tiempo de 20 minutos contenidos dentro de los intervalos de tiempo de 2 horas previos a los intervalos de tiempo a predecir. En este caso, los datos proporcionados corresponden a los días desde el 18 de Octubre de 2016 hasta el día 24 de Octubre de 2016 en los intervalos de tiempo de 6:00 a 8:00 y de 15:00 a 17:00. Los intervalos a predecir son de 8:00 a 10:00 y de 17:00 a 19:00 en esos mismos días.

Tabla *tabla_resultado_average_travel_time*

El esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
intersection_id	char(1)	Identificador de la intersección
tollgate_id	smallint	Identificador de la barrera de peaje
time_window	timestamp ARRAY[2]	Ventana de tiempo de 20 minutos
avg_travel_time	float	Tiempo medio de viaje (segundos)

Esta tabla contiene los intervalos que nos insta la competición a predecir en la primera fase de pruebas. Específicamente, en esta tabla se van a guardar las predicciones

realizadas con respecto al tiempo promedio de viaje en los intervalos a predecir (ver figura 1).

Tabla *tabla_resultado_traffic_volume*

El esquema de la tabla es el siguiente:

Campo	Tipo	Descripción
tollgate_id	smallint	Identificador de la intersección
time_window	timestamp ARRAY[2]	Ventana de tiempo de 20 minutos
direction	smallint	Dirección en la que se atraviesa la barrera de peaje
volume	int	Número de vehículos que atraviesan la barrera de peaje en la ventana de tiempo de 20 minutos

Esta tabla contiene los intervalos que nos insta la competición a predecir en la primera fase de pruebas. Específicamente, en esta tabla se van a guardar las predicciones realizadas con respecto al volumen de tráfico en los intervalos a predecir (ver figura 1).

COMPROBACIONES DE LOS DATOS

- Comprobar, mediante el script denominado *vehicle_trajectories_training_modified_foreign_key.sql*, que los identificadores que forman parte de cada uno de los objetos **link_object** que forman cada array **link_object[]** de la columna *travel_seq* de la tabla *vehicle_trajectories_training_modified* son válidos (es decir, existen en la tabla *road_links_modified*'.

- Comprobar que 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 que 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*.

BIBLIOGRAFÍA

- https://www.enterprisedb.com/downloads/postgres-postgresql-downloads (Página de descarga de PostgreSQL)