Consejo de diseño #172 Aprovecha tu modelo dimensional para análisis predictivo.

Análisis predictivo es el nombre que recibe una amplia gama de técnicas usadas para hacer predicciones sobre comportamientos futuros. Calificación de crédito, análisis de riesgo, y selección de promoción están entre las muchas aplicaciones que han probado ser útiles para generar ingresos y beneficios. Vale la pena echar un vistazo a la sección “análisis predictivo” de la Wikipedia para apreciar el amplio alcance de estas técnicas.

A pesar de las diferencias significantes entre las técnicas de análisis predictivo, casi todas pueden tomar los datos como series de “observaciones” normalmente codificadas por una clave de data warehouse conocida como cliente. Con motivo de la discusión asumiremos que tenemos 10 millones de clientes y esperamos predecir el comportamiento futuro poniendo en marcha una serie de aplicaciones analíticas predictivas sobre las historias de estos clientes.

Los datos de entrada deseados para la aplicación del análisis predictivo es una tabla de 10 millones de filas cuya clave primaria es la clave duradera y única del cliente, y cuyas columnas son un número abierto de atributos descriptivos, incluyendo los campos descriptivos usuales y los campos demográficos, *pero también incluyendo indicadores, cuentas y cantidades preparadas a propósito para la puesta en marcha de la aplicación de análisis predictivo.* Es este requisito en cursiva el que hace que esto sea tan interesante.

Nuestros datos de entrada al modelo de análisis predictivo parece una tabla muy amplia, pero es mucho más volátil y compleja que una tabla de dimensiones bastante estable. Algunos de los atributos ya existen en la dimensión normal de cliente, pero muchos de los atributos interesantes se eligen en el momento del análisis buscando en las tablas de hechos describiendo la historia de los clientes y después resumiendo esa historia en una etiqueta, puntuación o cantidad. Algunas aplicaciones de análisis predictivo pueden querer actualizar los datos de entrada en tiempo real, y el analista puede desear añadir de manera dinámica nuevos atributos calculados para los datos de entrada en cualquier momento.

Finalmente, una instantánea dada de estos datos de entrada puede ser una dimensión de cliente útil por propio derecho.

Si estas etiquetas de cliente, puntuaciones y cantidades son estables y se usan regularmente para filtrar consultas y agrupar más allá del análisis predictivo, entonces pueden llegar formar parte de la dimensión del consumidor de manera permanente. Dos patrones específicos de diseño hechos agregados como dimensiones de atributos y la etiqueta de comportamiento de las series de tiempo, se describen brevemente en el glosario de nuestra web e ilustrados en el capítulo 8 de la tercera edición de “The Data Warehouse Toolkit”.

Los datos de observación que alimentan una aplicación de análisis predictivo de un modelo dimensional es muy simple. Pero ¿cómo llenamos la tabla? ¿Se hace con el ETL mediante una compleja aplicación backroom que escribe esta tabla final en un formato convencional? ¿Se hace completamente con SQL en el momento de ejecución donde cada campo de cada columna se rellena con un comando SELECT correlacionado accediendo a tablas de hechos remotas? ¿Pueden los datos de los atributos de observación ser almacenados en una dimensión actual, antes que en tablas de hecho separadas, donde los múltiples valores se almacenan como un SQL ARRAY of STRUCTS, incluso como series de tiempo? ¿Hay herramientas de los entornos Hadoop como Spark o HBase que proporcionan un modo más natural y eficiente de construir el conjunto de datos de observación entregado para el análisis predictivo?

Parte de mi papel en este año de jubilación del Kimball Group es describir la relevancia duradera de nuestro bien-probado modelo de técnicas dimensionales, y al mismo tiempo desafiar a vosotros lectores de la comunidad data warehouse a pensar sobre las nuevas direcciones del data warehousing en el futuro. Para ser honesto, no tengo todas las respuestas a las preguntas mencionadas anteriormente, pero entre ahora y el final del año describiré al menos los desafíos y mi pensamiento de estos enfoques lo mejor que pueda.

<http://www.kimballgroup.com/2015/02/design-tip-172-leverage-dimensional-model-predictive-analytics/>