Diseño y Construcción de Data Warehouse
Curso 2017 - Proyecto

Análisis Multidimensional de Datos de la Red de monitoreo de la calidad del aire en Montevideo

1 Introducción

Hoy en día existen muchas organizaciones que promueven el mantenimiento de una buena calidad del Medio Ambiente. Es por esto que cada vez son más las autoridades que se sienten comprometidas en realizar distintos controles que permitan el monitoreo ambiental. En particular, la Intendencia de Montevideo ha implementado un programa de monitoreo de calidad del aire, que incluye una Red de Monitoreo que comenzó a operar en 2005.

El objetivo de la red es conocer la calidad del aire de base de Montevideo. Para este objetivo, la Red de Monitoreo opera con estaciones que permiten evaluar la concentración de material particulado en diversas fracciones y de gases que habitualmente se encuentran en ambientes urbanos (dióxido de azufre y dióxido de nitrógeno). Las estaciones de monitoreo están ubicadas en sitios fijos que permiten caracterizar la calidad de base de la ciudad. Las estaciones que funcionan actualmente en esta Red de Monitoreo están ubicadas en: Ciudad Vieja, Tres Cruces, Curva de Maroñas, Portones de Carrasco y Colón.

El juego de datos que se va a analizar presenta las mediciones históricas de las distintas estaciones para el período 2005 – 2015, así como información sobre el método de medida, lugar geográfico de las estaciones y sus modificaciones en el tiempo.

Para el trabajo planteado en este proyecto, también se cuenta con datos de las líneas de ómnibus de Montevideo y de los hogares que pertenecen a los distintos barrios de dicha ciudad.

Los datos de los ómnibus se obtienen de un archivo *Shapefile* que presenta los puntos geográficos de las ubicaciones de paradas de los recorridos de los ómnibus. Por otro lado, los datos de los hogares se obtienen de la Encuesta Continua de Hogares (ECH) realizada por el Instituto Nacional de Estadística (INE), que es un organismo gubernamental del Uruguay que se encarga de la realización y supervisión de las estadísticas nacionales, entre ellas la ECH. Todas las fuentes de datos antes mencionadas están disponibles en el Catálogo de Datos Abiertos de Uruguay.

En base a todos estos datos se busca analizar la polución que generan los ómnibus en Montevideo y cuánto afecta esta contaminación a los hogares de los distintos barrios de la ciudad. Para esto, se desea construir una plataforma que integre todas las fuentes de datos mencionadas y provea información confiable a través de una interfaz amigable y flexible orientada a la toma de decisiones.

2 Objetivos del proyecto

El objetivo principal de este proyecto es realizar un análisis multidimensional sobre algunos aspectos de la calidad del aire, de las líneas de ómnibus y de los hogares en la ciudad de Montevideo. En la Sección 3 se describen los requerimientos funcionales y no funcionales de la solución a desarrollar.

Para alcanzar este objetivo se deberá:

1- Realizar un **diseño conceptual multidimensional** de las dimensiones y relaciones dimensionales que surjan del análisis de los requerimientos.

2- Diseñar e implementar un **modelo lógico relacional** que dé soporte al modelo conceptual desarrollado en el punto 1, teniendo en cuenta las restricciones impuestas por las herramientas a utilizar.

3- Diseñar e implementar los **procesos de carga** del modelo lógico utilizando *Pentaho Data Integration* (también conocido como *Kettle*).

4- Implementar la solución completa de *Business Intelligence*.

3 Requerimientos

En esta sección se describen los requerimientos funcionales y no funcionales de la solución a desarrollar.

3.1 Requerimientos funcionales

Los Requerimientos 1 y 2 que se presentan a continuación, son los requerimientos funcionales que se deben satisfacer en el proyecto.

Requerimiento 1:

Se quiere analizar información acerca de la polución en el aire a través del tiempo. Las mediciones se desean analizar según las líneas de ómnibus, la ubicación geográfica, los métodos de monitoreo y los contaminantes en el aire.

Interesa visualizar: la cantidad de líneas de ómnibus que circulan en cada barrio, la polución y la cantidad de contaminantes en el aire en dichos barrios. Esta información debe poder agruparse por fecha, mes y año.

Las líneas de ómnibus se identifican con un número (por ejemplo línea 117). De los métodos de monitoreo interesa su nombre y una descripción del mismo, por ejemplo el método UYMVD_O3 mide el ozono (O3) con un sensor electroquímico. Por otro lado, sobre los contaminantes en el aire interesa su nombre, y se quiere poder clasificarlos según su subtipo y su tipo. Por ejemplo, el subtipo del dióxido de azufre (SO2) es compuesto de azufre (S) y éste corresponde al tipo contaminante primario. Finalmente, la ubicación geográfica se refiere a la estación donde se realiza el monitoreo del aire, la cual se quiere clasificar según el barrio en el que se encuentra.

Nota: En la sección *Fuentes de Datos* se presenta información acerca de los datos geográficos.

Requerimiento 2:

Se quiere analizar información acerca de los hogares afectados por la polución a lo largo del año 2013, teniendo en cuenta su ubicación geográfica y los contaminantes en el aire.

Los hogares se desean clasificar según el tipo de vivienda, la forma de tenencia, el nivel de confort, problemas de la vivienda y de acuerdo a su ubicación (si está o no en un asentamiento).

Interesa visualizar la cantidad de hogares afectados, así como también la cantidad de personas, cantidad de personas menores de 14 años, cantidad de personas mayores de 14 años, cantidad de hombres y cantidad de mujeres. Estas cantidades se quieren ver tanto sumadas como promediadas al agrupar según los distintos criterios.

Los indicadores antes mencionados se quieren visualizar a lo largo del año 2013 y se deben poder totalizar por fecha, día de la semana, mes, trimestre y semestre. Además, según la estación del año (otoño, invierno, primavera y verano).

Los tipos de viviendas son casa, apartamento y otros. Las formas de tenencia de vivienda son propietario, inquilino y ocupante. Los tipos de problemas de las viviendas pueden ser muros agrietados, goteras en el techo, etc.

Los elementos de confort son si cuenta con lavavajilla, secadora de ropa, microondas, etc. El nivel de confort de la vivienda se debe deducir a partir de la cantidad de electrodomésticos que tenga el hogar. Se debe clasificar en 4 niveles de confort, definiendo un rango de cantidad de electrodomésticos para cada nivel. También se puede tener en cuenta alguna otra información, como si tiene servicio doméstico o no, o el origen del agua que utiliza.

La ubicación geográfica se desea agrupar según el barrio y el centro comunal.

3.2 Requerimientos no funcionales

El principal requerimiento no funcional es que la solución deberá desarrollarse utilizando versiones estables de los productos del proyecto *Pentaho Business Intelligence Community* [1], en particular:

- *Pentaho BI Analytic*: la plataforma *Pentaho* se basa en una aplicación web J2EE que permite publicar y gestionar soluciones y un servidor que las implementa. Cada solución puede verse como una aplicación web que utiliza los diferentes servicios provistos por el servidor *Pentaho* (por ejemplo: motor OLAP, motor de workflow, servicios de data mining, etc.), y presenta la información al usuario mediante diferentes componentes (por ejemplo: reportes y gráficas dinámicas, vistas de análisis OLAP sobre cubos, tableros, diales con indicadores, etc.). Se sugiere utilizar la versión 7.0 de *Pentaho BI*, pero el estudiante podrá seleccionar la de su interés.

- *Pentaho Analysis Services (Mondrian)*: Este es un servidor OLAP del tipo ROLAP. Ya viene incluido en la instalación de la plataforma. También está disponible la herramienta *Schema Workbench* [1] que es una interface para definir esquemas en *Mondrian*.

- *Pentaho Data Integration (Kettle)*: Esta es la herramienta de ETL del proyecto *Pentaho*. Se sugiere utilizar la versión 7.0.

Dentro del Wiki de la edición *Community* de *Pentaho* pueden encontrarse documentación y tutoriales. En particular se sugiere documentación de *Pentaho BI Analytic* [2], de *Mondrian* [3] y de *Kettle* [4]. Otro material de apoyo puede ser encontrado en las referencias [5][6][7][8].

La plataforma utiliza RDBMs para almacenar la información del sistema (usuarios, roles, etc.) y para almacenar datos. Consulte la documentación de *Pentaho* [9] para saber cuáles son los RDBMs soportados y cómo se configura la conexión.

Para este proyecto se debe utilizar *PostgreSQL* [10], con su extensión Postgis, como RDBMs. Se sugiere *PgAdmin* [11] para gestionar de bases de datos *PostgreSQL*. No hay restricciones respecto al sistema operativo sobre el cual debe correr el prototipo, queda a elección del estudiante.

Además, se plantean los siguientes requerimientos adicionales:

1- En el desarrollo de la solución es importante tener en cuenta que existen varios componentes para presentar la información al usuario. Además de los componentes *JPivot* [12] y *Saiku* [13] (versión no *Enterprise*) que permiten hacer consultas OLAP y mostrar los resultados en forma tabular o mediante gráficas, como tarea opcional se podrá mostrar en un mapa de la ciudad de Montevideo alguno de los datos analizados. Para esto se sugiere estudiar las posibilidades de integración con *Google Maps* o de la utilización de *Saiku Chart Plus*.

2- Se pide implementar un reporte con *Pentaho Report Designer* [14] que esté publicado en la plataforma de *Pentaho*. Dentro del Wiki de la edición *Community* de *Pentaho* pueden encontrarse documentación y tutoriales [18].

3- Se pide de forma opcional, utilizando los componentes del *Community Dashboard Editor* (CDE) [15], construir un *Dashboard* en el cual se destaquen indicadores relevantes para la toma de decisiones a nivel gerencial. Ver ejemplos en *Webdetails* [16] de posibles implementaciones. Para un buen diseño de los *Dashboard* se sugiere aplicar conocimientos de HTLM, CSS y JavaScript. Además, componentes de la *Marketplace* [1] que faciliten el desarrollo, como *Bootstrap Dashboard Design* e *Ivy* [17].

4- La solución implementada deberá ser capaz de soportar nuevas cargas de datos sin sufrir modificaciones.

4 Fuentes de Datos

Para el proyecto se utilizara una base de datos *PostgreSQL* con su extensión Postgis, con información del monitoreo del aire, las líneas de ómnibus y los hogares de Montevideo, disponibles en Fuentes de datos internas y externas. Las mismas se listan a continuación:

Fuentes de datos internas:

- **Contaminantes:** Contiene la clasificación de los contaminantes monitoreados por la red de monitoreo [19].

Fuentes de datos externas:

- **Hogares**: Microdatos anonimizados de la Encuesta Continua de Hogares 2013 [20]. Se sugiere descargar el Diccionario 2013 donde se describen cada uno de los datos presentes en la Encuesta.

- **Líneas de ómnibus:** Shapefile de puntos con las ubicaciones de paradas de los recorridos de los ómnibus [21].
- **Barrios:** Shapefile con datos sobre los límites de los barrios de Montevideo [22].
- **Monitoreo de la calidad del aire:** En este juego de datos se presentan las mediciones históricas de las distintas estaciones para el período 2005 - 2015 así como información sobre el método de medida y lugar geográfico de las estaciones y sus modificaciones en el tiempo [23].

De las estaciones de monitoreo del aire interesa su geometría, que en este caso es un multipolígono denominada *geom*. De las líneas de ómnibus también interesa su geometría, pero en este caso es un punto y también se denomina *geom*.

Para conocer los puntos que pertenecen a un área geográfica es necesario usar la siguiente función de Postgis: *ST_Within(A, B)*, esta verifica si el punto A pertenece al área B [24].

5 Resultados esperados

Al finalizar este proyecto se espera contar con:

1. Un prototipo funcional que abarque los requerimientos funcionales y no funcionales planteados en la Sección 3.

2. Un informe que describa la solución propuesta. Este documento deberá incluir al menos:

    a. Análisis de requerimientos.
    b. Diseño conceptual de la solución, argumentando las decisiones de diseño tomadas.
    c. Diseño lógico de la solución, argumentando las decisiones de diseño tomadas.
    d. Implementación de las relaciones dimensionales y dimensiones sobre *Pentaho BI Server* (archivos .xml generados)
    e. Documentación sobre el proceso de carga (fuentes consideradas y pseudocódigo del proceso de carga)
    f. Documentación sobre todos los componentes que se incluyeron en la solución para satisfacer los requerimientos.
    g. Descripción de los problemas de calidad de datos encontrados y planteo de las soluciones propuestas para los mismos.
    h. Esbozo de un plan de testeo de la solución.

Al finalizar el proyecto se realizará una defensa del mismo y se entregará: una versión impresa del informe junto con un CD conteniendo el prototipo implementado y la documentación.

Referencias

[1] http://community.pentaho.com/
[2] http://wiki.pentaho.com/display/COM/Community+Wiki+Home
[3]http://wiki.pentaho.com/display/analysis/Pentaho+Analysis+Community+Documentation
[4]http://wiki.pentaho.com/display/EAI/Pentaho+Data+Integration+%28Kettle%29+Tutorial
[5] http://pedroalves-bi.blogspot.com.uy/
[6] http://pentaho-bi-suite.blogspot.com.uy/
[7] http://forums.pentaho.com/
[8] http://pentaho.almacen-datos.com/tutorial.html
[9] http://wiki.pentaho.com/display/ServerDoc2x/.04+Configuring+Data+Sources
[10] https://www.postgresql.org/download/
[11] https://www.pgadmin.org/download/
[12]http://wiki.pentaho.com/display/ServerDoc1x/Introduction+to+Pentaho+Analysis+Viewer+(JPivot)
[13] http://community.meteorite.bi/
[14] http://community.pentaho.com/projects/reporting/
[15] http://community.pentaho.com/ctools/cde/
[16] http://www.webdetails.pt/
[17] http://ivy-is.co.uk/ivy-labs/pentaho-bootstrap-dashboards/
[18]http://wiki.pentaho.com/display/Reporting/Pentaho+Reporting+-+User+Guide+for+Report+Designer
[19] https://eva.fing.edu.uy/course/view.php?id=648#section-8
[20] http://www.ine.gub.uy/encuesta-continua-de-hogares1
[21]https://catalogodatos.gub.uy/dataset/transporte-colectivo-paradas-y-puntos-de-control
[22] https://catalogodatos.gub.uy/dataset/limites-barrios/resource/bc509d2e-f5df-44fd-86af-3db9f3bb071c
[23]https://catalogodatos.gub.uy/dataset/red-de-monitoreo-de-la-calidad-del-aire-de-montevideo
[24] http://www.postgis.org/documentation/manual-svn/ST_DWithin.html