Diseño y Construcción de Data Warehouse
Curso 2018 - Proyecto

Análisis multidimensional de datos de los contenedores distribuidos en Montevideo

1 Introducción

Hoy en día existen distintos tipos de contenedores distribuidos en toda la ciudad de Montevideo:

- contenedores de residuos: estos son los contenedores en los cuales los ciudadanos depositan la basura de sus domicilios.
- contenedores de residuos reciclables: estos están habilitados para el depósito de residuos secos. Este tipo de contenedor está ubicado en supermercados, complejos habitacionales, edificios, vía pública, etc.
- contenedores de materiales especiales: estos contenedores contienen plásticos, latas, vidrios y pilas.

Para el trabajo planteado en este proyecto, se cuenta con datos provistos por la Intendencia de Montevideo. Dichos datos refieren a la ubicación de los contenedores de residuos domiciliarios y a los circuitos de recolección a los que éstos pertenecen. También, se cuenta con información acerca de la ubicación de contenedores de residuos reciclables y materiales especiales.

Por otro lado, existe información acerca de las empresas alimenticias y hogares de la ciudad de Montevideo. La información de las empresas también es provista por la Intendencia de Montevideo, mientras que 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.

Para el trabajo planteado en este proyecto, también se cuenta con datos de los distintos barrios, centros comunales zonales (CCZ) y municipios de la ciudad de Montevideo. Estos datos, también publicados por la Intendencia de Montevideo, se obtienen de archivos *Shapefile* que presentan los puntos geográficos que determinan los límites de los espacios geográficos antes mencionados.

Teniendo en cuenta distintos perfiles de personas u organizaciones interesados en analizar los hogares y empresas alimenticias que son afectados por los contenedores de residuos distribuidos en la ciudad de Montevideo, se desea construir una plataforma que integre todas las fuentes de datos antes mencionadas. Dicha plataforma debe proveer

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 los hogares y empresas alimenticias que son afectados por los contenedores distribuidos en los diferentes barrios de 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 de la recolección de basura de los contenedores de acuerdo al turno y lugar en el cual se hace la recolección.

Interesa visualizar la frecuencia semanal y la cantidad de contenedores de basura por circuito de recolección, por municipio, por turno semanal y por turno horario.

Los indicadores antes mencionados se quieren visualizar según los contenedores de residuos, su ubicación geográfica, el turno horario y el turno semanal en el cual se hace la recolección de dichos contenedores.

El turno semanal refiere al día de la semana en el cual se hace la recolección y el turno horario refiere a la hora en la que se hace la recolección. En el caso de los días, interesa clasificarlos según el tipo de día (semanal o fin de semana). Mientras que en el caso de las horas, interesa agruparlas por rangos y a éstos por tipo de rango (matutino,

vespertino o nocturno). Por ejemplo, un contenedor de basura se recolecta los lunes, miércoles y viernes de 14 a 16 hs. en el horario vespertino.

Respecto a los contenedores de residuos se conoce su identificador y su ubicación (coordenadas x, y). Es importante mostrar la información clasificando a los contenedores según el circuito de recolección al cual pertenecen. Del circuito se conoce un identificador y los vértices del polígono que determinan los límites del mismo. A su vez, los circuitos se agrupan por municipio. Estos contenedores también se agrupan por barrios, éstos por CCZ y luego por municipios.

De los contenedores de residuos reciclables se conoce identificador, coordenadas (x, y) de su ubicación y nombre y dirección del local en el cual están ubicados. Éstos, son clasificados según el tipo de local en el cual se encuentra (edificio, supermercado, etc.).

De los contenedores de materiales especiales se conoce su identificador y la dirección en la cual están ubicados. Estos contenedores se clasifican según el material que contienen (pilas, plástico, vidrio, etc.).

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 y empresas alimenticias que son afectados por los contenedores de basura.

Interesa visualizar la cantidad de empresas y la cantidad de hogares, así como también la cantidad de personas, cantidad de personas menores de 14 años y cantidad de personas mayores de 14 años. Estas cantidades se quieren ver tanto sumadas como promediadas al agrupar según los distintos criterios. Por otro lado, interesa visualizar los siguientes índices: cantidad de contenedores por personas y cantidad de contenedores por hogares. Estos índices deben ser calculados como se muestra a continuación:

(Cantidad de contenedores / cantidad de personas) x 100

(Cantidad de contenedores / cantidad de hogares) x 100

Los indicadores antes mencionados se quieren visualizar según los contenedores de residuos, su ubicación geográfica, los hogares y las empresas alimenticias.

Los hogares tienen un identificador y se clasifican según el tipo de vivienda, los problemas de la misma y de acuerdo a su estrato social. Además, interesa la clasificación de hogares según cuenta con sanitaria o no y si está en un asentamiento o no. Los tipos de viviendas son casa, apartamento y otros. Los tipos de problemas de las viviendas pueden ser humedad en techo, poca luz solar, etc. El estrato social puede ser nivel económico alto, bajo, etc.

De las empresas alimenticias se conoce su RUT, su razón social y dirección. Dichas empresas son clasificadas por tipo (elaboradores, expendedores o elaboradores-expendedores) y de acuerdo al estado de habilitación (habilitado, en trámite o vencido).

Finalmente, la ubicación geográfica se desea agrupar según el barrio, el CCZ y el municipio. Tanto para los barrios, los CCZ y los municipios se conoce su identificador y su nombre.

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* [1], en particular:

- *Pentaho BI Analytic* [2]: 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 8.0 de *Pentaho BI*, pero el estudiante podrá seleccionar la de su interés, teniendo en cuenta que se deben obtener todos los resultados esperados.

- *Pentaho Analysis Services, Mondrian* [3]: Este es un servidor OLAP del tipo ROLAP. Además, se cuenta con la herramienta *Schema Workbench* [1] que permite definir esquemas en *Mondrian* y luego publicarlos. Estas herramientas están incluidas en la instalación de la plataforma BI.

- *Pentaho Data Integration, Kettle* [4]: Esta es la herramienta de ETL del proyecto *Pentaho*. Se sugiere utilizar la versión 8.0 que es la incluida en la instalación de la plataforma BI.

- *GeoKettle* [5]: Es una versión de la herramienta genérica *Kettle*, para el manejo de datos espaciales. Se sugiere utilizar la versión 2.5.

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]. Por otro lado, se encuentra *GeoKettle* [5], que es una herramienta para la integración de fuentes de datos espaciales.
En las referencias [6][7][8] puede ser consultado material de apoyo para el uso de estas herramientas.

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 sugiere utilizar *PostgreSQL10* [10], con su extensión *Postgis* [11], como RDBMs. Se sugiere *PgAdmin4* [12] 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:

- 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* [13] y *Saiku* [14] (versión no *Enterprise*) que permiten hacer consultas OLAP y mostrar los resultados en forma tabular o mediante gráficas, como tarea opcional investigar la visualización 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 la utilización del *plugin Saiku Chart Plus* [15].

- Se pide implementar un reporte con *Pentaho Report Designer* [16] que esté publicado en la plataforma de *Pentaho*. Esta herramienta está incluida en la instalación de la plataforma BI.

- Se pide, utilizando los componentes del *Community Dashboard Editor* (CDE) [17], construir uno o varios *Dashboards* en los cuales se destaquen indicadores relevantes para la toma de decisiones a nivel gerencial. Ver ejemplos en *Webdetails* [18] de posibles implementaciones. Para un buen diseño de los *Dashboards* se sugiere aplicar conocimientos de HTLM, CSS y JavaScript. Además, componentes de *Pentaho Marketplace* [15] que faciliten el desarrollo de los mismos. *Pentaho Marketplace* ofrece a los usuarios diferentes *plugins*. En particular, *Ivy Bootstrap* es un *plugin* de *Pentaho* que contiene una selección de componentes personalizados para la creación de *Dashboards* dinámicos, un ejemplo puede observarse en *Bootstrap Dashboard Design* e *Ivy* [19].

4 Fuentes de Datos

Las fuentes de datos necesarias para la realización del proyecto son las que se listan a continuación:

- **Hogares**: Microdatos anonimizados de la Encuesta Continua de Hogares 2016 [20]. Se sugiere descargar el Diccionario 2016 donde se describen cada uno de los datos presentes en la Encuesta.
- **Empresas alimenticias:** El Servicio de Regulación Alimenticia realiza el registro y habilitación de distintas empresas. Información de las habilitaciones se encuentra en [21].
- **Barrios:** *Shapefile* con datos sobre los límites de los barrios de Montevideo, según la división del Instituto Nacional de Estadística (INE) [22].
- **Centros comunales zonales:** *Shapefile* de polígonos de los 18 centros comunales zonales de la ciudad de Montevideo [23].
- **Municipios:** *Shapefile* de polígonos con los límites de los Municipios de la ciudad de Montevideo [24]
- **Contenedores de residuos domiciliarios:** Ubicación de contenedores de residuos domiciliarios con la frecuencia de recolección programada, y los circuitos de recolección a los que pertenecen [25].
- **Contenedores de residuos reciclables:** Ubicación de contenedores de residuos secos [26].
- **Contenedores de materiales especiales:** *Shapefile* con información sobre la ubicación de los contenedores especiales [27].

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. Tener en cuenta que **no deben considerarse** los requerimientos correspondientes a los contenedores de residuos reciclables ni a los contenedores de materiales especiales.

2. Un informe que describa la solución propuesta. Este documento deberá incluir al menos:
    a. Análisis de requerimientos.
    b. Diseño conceptual completo de la solución, argumentando las decisiones de diseño tomadas.
    c. Diseño lógico completo de la solución, argumentando las decisiones de diseño tomadas.
    d. Implementación de las relaciones dimensionales y dimensiones sobre *Pentaho BI Server* (archivos .xml generados). En el caso de las dimensiones **no deben considerarse** las correspondientes a los contenedores de residuos reciclables ni a los contenedores de materiales especiales.
    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. Análisis de la capacidad que presenta, la solución implementada, de soportar nuevas cargas de datos.
    h. Descripción de los problemas de calidad de datos encontrados y planteo de las soluciones propuestas para los mismos.
    i. 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://www.pentaho.com/product/version-8-0
[2] https://wiki.pentaho.com/display/ServerDoc2x
[3] https://wiki.pentaho.com/display/analysis/Pentaho+Analysis+Community+Documentation
[4] https://wiki.pentaho.com/display/EAIes
[5] http://www.spatialytics.org/projects/geokettle/#
[6] http://forums.pentaho.com/
[7] http://pentaho.almacen-datos.com/tutorial.html
[8] http://www.spatialytics.org/files/foss4g2011/geokettle-workshop.pdf
[9] https://help.pentaho.com/Documentation/7.1/0H0/Specify_Data_Connections_for_the_Pentaho_Server/Define_Data_Connections_for_the_Pentaho_Server
[10] https://www.postgresql.org/
[11] https://postgis.net/docs/manual-dev/postgis-es.html
[12] https://www.pgadmin.org/
[13] https://sourceforge.net/projects/jpivot/?source=navbar
[14] http://community.meteorite.bi/
[15] http://www.pentaho.com/marketplace/
[16] http://community.pentaho.com/projects/reporting/
[17] http://community.pentaho.com/ctools/cde/
[18] http://www.webdetails.pt/
[19] http://ivy-is.co.uk/ivy-labs/pentaho-bootstrap-dashboards/
[20] http://ine.gub.uy/web/guest/encuesta-continua-de-hogares1
[21] https://catalogodatos.gub.uy/dataset/habilitacion-registro-de-locales-y-empresas-alimentarias
[22] https://catalogodatos.gub.uy/dataset/limites-barrios
[23] https://catalogodatos.gub.uy/dataset/limites-centros-comunales-zonales
[24] https://catalogodatos.gub.uy/dataset/limites-de-municipios-de-montevideo
[25] https://catalogodatos.gub.uy/dataset/contenedores-de-residuos-domiciliarios-ubicacion-circuitos-y-frecuencia-de-recoleccion-programada
[26] https://catalogodatos.gub.uy/dataset/contenedores-residuos-secos-domicialiarios
[27] https://catalogodatos.gub.uy/dataset/contenedores_reciclable