Almacén de datos espacial para el análisis de desastres naturales en el continente americano

Josefa Somodevilla[1] , Yuridiana Alemán[1], Helena Gómez[1], Nahun Loya[1],
[1] FCC, Benemérita Universidad Autónoma de Puebla, México
{mariajsomodevilla, yuridiana.aleman, helena.adorno, nahun.loya }@gmail.com
(Paper received on August 10, 2012, accepted on August 24, 2012)

Resumen. Este documento presenta un sistema para soporte de toma de decisiones usando el dominio de los desastres naturales. Partimos de diversos conjuntos de datos que son obtenidos de distintos repositorios, a los cuales se les aplica técnicas de pre-procesamiento para tener un conjunto de datos confiable. Además incorporamos datos espaciales, por ejemplo la superficie de los países, vistos como un dato geométrico dentro de la base de datos. Se propone un modelo de estrella multidimensional usando el motor de base de datos SQL SERVER 2008 para representar el modelo.

Palabras clave: Desastres naturales, Sistema para la toma de decisiones, bases de datos geográficas, almacén de datos.

1 Introducción

Un desastre natural es un evento que produce daños nocivos en la economía, el medio ambiente y/o en la salud de la población. Los desastres naturales son fenómenos que se presentan de diferentes formas y en distintas épocas del año, por lo tanto, el estudio de los desastres naturales es importante ya que estos provocan la pérdida de vidas humanas, hay daños inconmensurables en la estructura de las ciudades, servicios y en la ecología. El principal problema cuando ocurre algún tipo de desastre es que no se tiene la información correcta para enfrentarlo, desde que se presenta hasta que concluye, por tanto es necesario contar con diversos mecanismos para la toma de decisiones haciendo uso de la tecnología. Diversos autores han usado los modelos de toma de decisiones para prevención de eventos de destrucción masiva, por ejemplo en [1] proponen un sistema de toma de decisiones para prevenir riesgos por sismos, usando los datos de U.S Geological Survey y Sistema Sísmico Nacional, ellos incorporan operaciones OLAP y usan minería de datos con el objetivo de encontrar patrones para predecir el comportamiento de los sismos. Los sistemas de toma de decisiones son importantes en diversos dominios, tal es el caso de [2], quienes proponen usar la tecnología de Datawarehouse y los sistemas de información geográfica para pronosticar las mejores áreas de cultivo en las zonas del Brasil, ellos incorporan datos espaciales para representar: municipalidades, zonas y regiones de cultivo, finalmente proponen una forma adecuada de ejecutar las consultas considerando los datos espaciales.

Special Issue in Information Technology
Research in Computing Science 57, 2012, pp. 39-45

En este trabajo se realiza una recolección de datos correspondientes a los desastres que han ocurrido desde el año 1980 hasta 2011. El objetivo es que la información presentada como resultado sirva de soporte para la toma de decisiones en la prevención y mitigación de desastres, en particular de los países del continente americano. Los datos son obtenidos de *The international disaster database* (EM-DAT) [3] y se les aplica el proceso KDD (*Knowledge Discovery in Databases)* con el objetivo de extraer conocimiento útil que sirva de apoyo para la toma de decisiones. El estudio se enfoca en un número finito de desastres naturales: sequías, terremotos, epidemias, temperaturas extremas, inundaciones, tormentas, erupciones volcánicas e incendios forestales. Se incorpora el uso de los datos espaciales correspondientes a los países de América y son representados en el marco de la tecnología de SQL server, de esta manera se pueden plantear cuestionamientos relacionados con operaciones topológicas entre los países y los desastres naturales.

2 Descripción de los conjuntos de datos

Las fuentes de información consideradas en este estudio son proporcionadas por EM-DAT, los datos espaciales se obtienen manualmente a través de la herramienta Google Earth [4], los datos referentes a las diversas acepciones de los nombres de países y ciudades son obtenidos a través del repositorio en línea Geonames [5] y finalmente los datos del Banco Mundial (BM) [6].

- *EM-DAT* desde 1988 es mantenida por el *Centro de Investigación sobre la Epidemiología de los Desastres (CRED).* Contiene datos relevantes acerca de 18,000 desastres que han ocurrido desde el año 1900. Esta base de datos es compilada de diversas fuentes incluyendo agencias de la Organización Naciones Unidas (ONU), organizaciones no gubernamentales, compañías de seguros, institutos de investigación y agencias de noticias. Además, proporciona datos de diferentes grupos de desastres entre los que podemos destacar: biológicos, climatológicos, geofísicos, hidrológicos, meteorológicos y tecnológicos.
- *Google Earth* comprende un conjunto de herramientas para la exploración visual de datos, haciendo uso de esta tecnología se obtuvieron los polígonos de cada país, esto para establecer operaciones topológicas de todo el sistema de toma de decisiones en particular con los datos espaciales.
- *Geonames* es un repositorio de datos en línea que provee información acerca de nombres de países y sus posibles acepciones, población, capitales, zonas horarias, código postal, código ISO, niveles de elevación, ubicación de los países en base a su latitud y longitud.
- *BM* es una institución internacional que conforma distintas bases de datos de temas diversos, nosotros usamos los datos referentes a población actual y el Producto Interno Bruto (PIB) de los países.

El conjunto de datos a considerar puede ser resumido en la Tabla 1, en la cual se observa que el mayor número de atributos se obtiene a partir de la base de datos EM-DAT.

Tabla 1. Conjunto de datos y atributos considerados.

Organización	No de registros	No. de Atributos	Atributos
EM-DAT	2429	12	año, clave, dis_subgrupo, dis_tipo, fecha_ini, fecha_fin, no_killed, no_heridos, no_afectados, no_sincasa, total_afectados, total_daño
GOOGLE EARTH	45	3	Longitud, latitud, polígono.
GEONAMES	2429	2	País_acep, nombre_pais, Ciudad_acep.
BM	2429	2	Población, PIB.

Una forma práctica de mostrar el conjunto de datos a trabajar es de forma gráfica. En la Figura 1 se muestra el comportamiento de los desastres a través de tiempo, la gráfica refleja el número de desastres del periodo comprendido entre 1980 y 2011. Por otra parte en la Figura 2 se presentan los diferentes tipos de desastres considerados en el estudio. Se puede notar que los de tipo meteorológico e hidrológico son los que ocurren con mayor frecuencia en el conjunto de datos

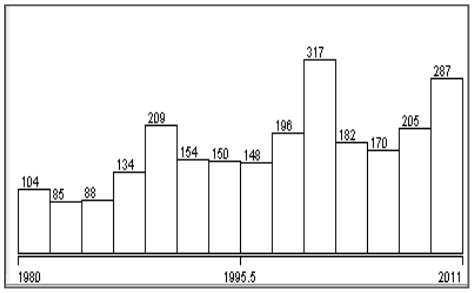

Figura 1. Número total de desastres naturales en el periodo 1980-2011

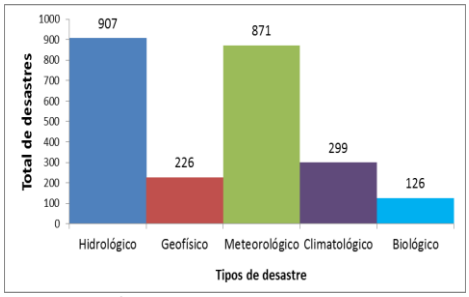

Figura 2. Número de desastres naturales de acuerdo al subgrupo de 1980 hasta 2011.

Para el caso de estudio se consideró todo el continente americano, donde la mayor parte de los desastres son en Estados Unidos (665) y México (182). Sin embargo por parte de Sudamérica se observa que en países como Brasil, Colombia y Perú también se han presentado eventos de destrucción masiva que han provocado diversos daños en la población y en los ecosistemas en general, con un total de 394 entre estos tres países.

3 Diseño del sistema para la toma de decisiones.

Los Almacenes de Datos y operaciones OLAP son de importante utilidad para analizar grandes cantidades de datos. Estos datos que por lo general son extraídos de bases de datos transaccionales, que frecuentemente contienen información espacial, la cual resulta muy útil para el proceso de toma de decisiones [7].

En este trabajo proponemos un sistema de soporte para el análisis de los desastres naturales ocurridos en el continente americano. A partir de la información recolectada, se diseñó en primer lugar un modelo conceptual de la base de datos de

desastres naturales, el cual se presenta en la Figura 4 mediante un diagrama Entidad Relación Extendido (ER), con pictogramas para representar conceptos espaciales [7]. El segundo paso fue realizar el mapeo del modelo ER a un modelo relacional multidimensional, que posteriormente fue implementado en una base de datos SQL Server. La arquitectura del sistema se presenta en la Figura 3.

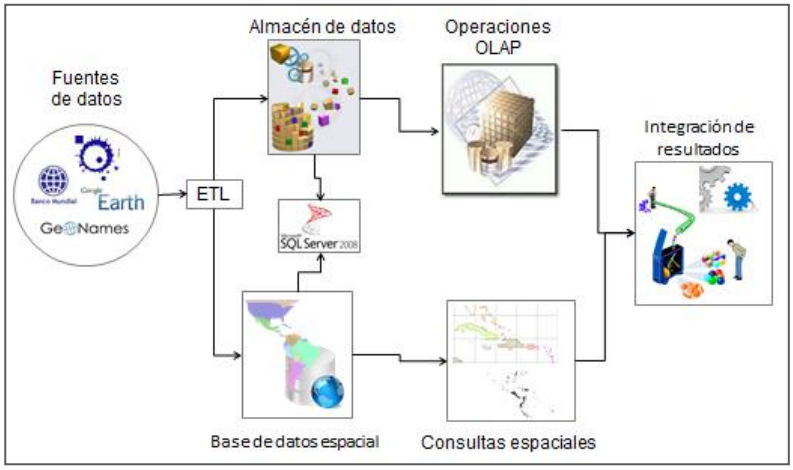

Figura 3. Arquitectura del Sistema para soporte a la toma de decisiones en el dominio de Desastres Naturales.

3.1 Proceso de ETL

La información fue recopilada de diferentes fuentes, en este caso de los diferentes repositorios y bases de datos. Se realizó un proceso de ETL, siglas que en inglés significan Extracción, Transformación y Carga (Figura 3). Dicho proceso se detalla a continuación:

Extracción: Se obtienen los datos de EM-DAT, GEONAMES, BM y Google Earth, y se analizan como se muestra en la Tabla 1. Se integran los datos de las diferentes fuentes y se obtiene un nuevo conjunto de datos listo para la creación del almacén de datos. Se estandarizaron los nombres de las ciudades de los datos obtenidos en EM-DAT con GEONAMES de acuerdo con sus diferentes acepciones. Para los casos que no se pudieron empatar las ciudades automáticamente se realizó una limpieza manual, revisando errores de ortografía, e insertando manualmente aquellas ubicaciones que no estaban contempladas en la base de datos de GEONAMES.

Transformación: Los valores obtenidos a través de la herramienta Google Earth, es decir, las latitudes y longitudes de los bordes de cada país se América fueron transformados en coordenadas geométricas que posteriormente fueron insertadas en forma de polígono, para tener una representación adecuada en la herramienta de desarrollo. También se realizó un control sobre las ubicaciones geográficas de cada ciudad, verificando que cada punto este contenido en su país correspondiente.

Encontramos muchas ciudades que estaban situadas en países que no correspondían y también ciudades situadas en el mar, todos estos casos fueron resueltos manualmente.
Carga (Load): Los datos transformados, son cargados en el sistema utilizando tecnología de SQL Server para almacenes de datos.

3.2 Esquema Multidimensional

Para el diseño del almacén de datos se usó la metodología de Kimball [8,9] basado en un modelado multidimensional en el cual se utilizó un modelo estrella como se muestra en la Figura 5, donde la tabla de hechos contiene la información relevante de cada desastre. Los datos que se consideran son el número de muertos, afectados, heridos, la pérdida económica y el nombre del desastre (aunque este atributo no existe para todos los registros).

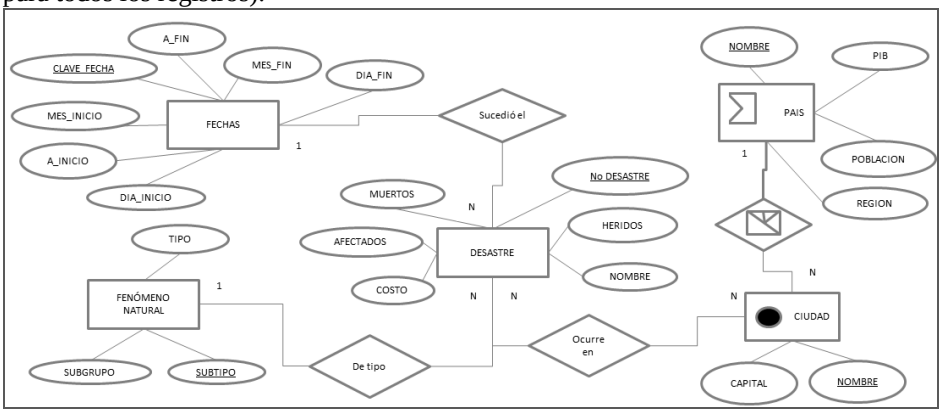

Figura 4. Modelo Conceptual ER extendido con conceptos espaciales, de base de datos para desastres naturales

Se consideran tres dimensiones para la construcción del cubo de datos:

- Dim_Fechas: Considerada la dimensión temporal, aquí se especifica la fecha exacta del inicio y fin de cada desastre
- Dim_Localidad: Esta es la dimensión espacial, aquí se considera principalmente la ciudad en la que ocurrió el desastre y el país al que pertenece dicha ciudad. Como atributos espaciales, para cada ciudad de registra su ubicación en forma de un par de coordenadas (punto) y la forma geométrica del país (polígono).
- Dim_Fenómeno_Natural: Esta dimensión almacena las principales características de cada desastre, en cuanto a su clasificación en subgrupos y subtipos, los subtipos se describen en la Figura 2.

3.3 Operaciones OLAP

Las herramientas OLAP permiten al usuario obtener una visión multidimensional de los datos, se pueden realizar consultas sin tener conocimiento de la estructura interna del almacén de datos. Algunos resultados interesantes obtenidos mediante

consultas OLAP se muestran en la Figura 6, en la cual se presenta el total de desastres agrupados por Continente, América Central (CAM), América del Norte (NAM), América del SUR (SAM), el Caribe, y subgrupo de desastre, excluyendo al año 2012.

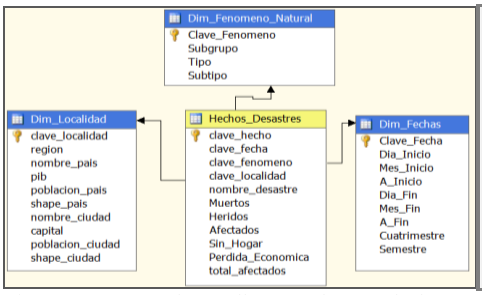

Figura 5. Esquema de estrella para Almacén de datos.

Año					
Excluding: 2012					
	Region				
	CAM	Caribe	NAM	SAM	Grand Total
Subgrupo	Total	Total	Total	Total	Total
Biological	34	14	13	65	126
Climatological	36	22	133	107	298
Geophysical	57	14	54	101	226
Hydrological	139	109	227	431	906
Meteorological	78	225	507	61	871
Grand Total	344	384	934	765	2427

Figura 6. Consulta OLAP.

4 Conclusiones y Trabajo Futuro

En este trabajo hemos presentado un sistema para la toma de decisiones de los desastres naturales, mediante el cual podemos analizar la información histórica, con el objetivo de lograr una mitigación de fenómenos de destrucción masiva. Se pueden visualizar las zonas de riesgo más inminentes, en referencia a cada fenómeno de destrucción masiva considerado en este trabajo. También se pueden estudiar los desastres naturales desde la perspectiva geográfica y el costo monetario generado por tales hechos. Esta herramienta puede ayudar a reducir el tiempo en la toma de decisiones, automatizando las tareas de administración y valoración de un modelo toma de decisiones, especialmente importante para usuarios que no tienen conocimiento profundo acerca de la teoría computacional. Actualmente se está trabajando en la integración de herramientas de minería de datos para la explotación del almacén de datos propuesto y de esta manera generar conocimiento para dar soporte a la toma de decisiones.

5 Referencias

[1]. Somodevilla, J., Priego, A., Castillo, E., Pineda, I., Vilariño, D., Nava, A.: "Decision support system for seismic risks". Journal of Computer Science and Technology, Vol. 12, No. 2, Argentina, (2012)

[2]. Sampaio, M.C., de Sousa, A.G., Baptista, C.d.S.: Towards a logical multidimensional model for spatial data warehousing and olap. In: Proceedings of the 9th ACM international workshop on Data warehousing and OLAP. DOLAP'06, New York, NY, USA, ACM (2006) 83-90

[3]. Université Catholique de Louvain - Brussels – Belgium "EM-DAT: The OFDA/CRED International Disaster Database" www.emdat.be

[4]. Keyhole, Inc. y Google ww.earth.google.com, (2005).

[5]. Geonames, Creative Commons, www.geonames.org, Online access may–15-12 (2012).
[6]. The World Bank Group, Online access may–15-12, (2012).
[7]. Shekhar, S., Chawla, S.: "Spatial Databases: A Tour" Prentice Hall,(2003)
[8]. Kimbal, R., Ross, M., Thornthwaite, W., Mundy, J., Becker B.: Relentlessly Practical Tools for Data Warehousing and Business Intelligence, The Kimball Group Reader (2010).
[9]. Kimball, R., Reeves, L., Ross, M., Thornthwaite, W.: The Data Warehouse Lifecycle Toolkit. 2nd Edition. New York, Wiley, (2008).