BENEMÉRITA
UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

MAESTRIA EN CIENCIAS DE LA COMPUTACIÓN

TESIS

Modelo para la Generación y Consulta de un Cubo de Datos Espacial Difuso

PRESENTA:

DAVID PÉREZ GÓMEZ

24 DE OCTUBRE DEL 2007

INDICE

INTRODUCCIÓN

En la actualidad los seres humanos tienen una idea ya formada de los panoramas que la naturaleza les presenta, cuando se mira alrededor de los paisajes las colecciones de objetos que nuestros ojos alcanzan a ver, son transformadas por nuestro cerebro en un sin fin de formas, nosotros las relacionamos gracias a experiencias y entrenamientos que nos permiten construir modelos y reconocer patrones que han sido esenciales para el desarrollo del humano [1].

Uno de los enfoques que se le ha dado a la representación de los modelos naturales es la relación entre un mapa cartográfico y un sistema de información [2]. Los Sistemas de Información Geográfica (GIS), son usados por un amplio rango de disciplinas tanto técnicas como académicas [1], ya que los GIS proporcionan herramientas universales para el manejo de datos espaciales, un dato espacial referencia una coordenada o ubicación real en la tierra.

Otra herramienta que ha surgido como apoyo al enfoque de acercar los hechos del mundo real a un modelo formal es la lógica difusa. La lógica booleana está muy restringida en poder llevar la percepción humana a un modelo matemático, ya que está basada en dos valores (0 y 1), que significan el pertenecer o no a un conjunto. Sin embargo, la lógica difusa permite generar una clasificación más cercana a la percepción humana, ya que un objeto puede pertenecer a varias clases al mismo tiempo, con diferentes grados de pertenencia, con ello se enriquece el valor semántico que pueda tener dicho objeto.

Para poder integrar las áreas de la lógica difusa con los datos espaciales que nos proporcionan los GIS, se debe contar con una estructura que permita plasmar toda la información que dichas áreas produzcan y que al mismo tiempo simplifiquen y agilicen las consultas a grandes volúmenes de información.

Una solución al problema descrito en el párrafo anterior podría ser el uso de un almacén de datos, este tipo de estructuras son capaces de almacenar grandes volúmenes de información, como lo pueden ser los datos espaciales, otra ventaja de este tipo de estructura de datos, es el futuro uso que pueden hacer de él diversas herramientas en el análisis de datos, por ejemplo: la minería de datos, hace uso de un almacén para la extracción de información mediante diversas técnicas que ayudan en el proceso de generación del conocimiento, las herramientas de análisis de procesamiento analítico en línea (OnLine Analytical Processing - OLAP), muy usado en el ámbito de Inteligencia en los Negocios (Bussines Intelligence -BI), mediante la generación de un cubo de datos el cual represente la información de una manera multidimensional relacionando diversos temas que son denominados dimensiones.

La Tesis está organizada en 4 capítulos, el capítulo 1 explica cada uno de los conceptos que a lo largo de este trabajo vamos a utilizar, el capítulo 2 presenta el Análisis y el Diseño de el Almacén de Datos Espacial Difuso, en este capítulo se explicará el modelo a seguir en la elaboración del almacén y la integración de la lógica difusa con las bases de datos espaciales. En el capítulo 3, se realiza la Implementación del Cubo de Datos mediante las diversas herramientas que nos permiten representarlo, finalmente el capítulo 4 presenta el caso de estudio propuesto, con el fin de poder representar un cubo

de datos con la información que ha sido recopilada del almacén toda vez que éste ha sido construido.

CAPÍTULO 1

MARCO TEÓRICO

1.1 Sistemas de Información Geográfica

La definición de Sistema de Información Geográfica establece que un GIS es "un sistema computarizado para capturar, manejar, integrar, manipular, analizar y desplegar datos que están espacialmente referenciados a la tierra" [3][4][5]. Como se mencionó anteriormente, un dato espacial es la información de una coordenada o coordenadas de un punto(s) que hace referencia a una ubicación real en la tierra, así pues los GIS están ligados con las bases de datos espaciales, ya que la información contenida en la base de datos será transformada para ser representada en forma digital por un GIS.

Un GIS moderno hace posible integrar la información que es difícil de asociar a través de algún otro medio y combina variables mapeadas para construir y analizar nuevas variables [2].

Un GIS posee varias funcionalidades entre las que se pueden mencionar: la capacidad para manejar varias capas temáticas de datos y relacionarlas para los mismos puntos en el espacio, combinar diferentes capas para obtener una nueva capa temática con la información de las diferentes capas y hacer un mapeo de los resultados obtenidos de un análisis que se realice. Dado que se está trabajando con información georeferenciada, esta debe de tener una localización dentro de un sistema de coordenadas, como base de referencia. La representación más común de información espacial es un mapa donde la localización de un punto por ejemplo, puede ser dada haciendo uso de la latitud y la longitud.

Los GIS, pueden ser definidos de forma general, como un conjunto de técnicas empleadas para alcanzar los siguientes objetivos:

- Encontrar las localizaciones apropiadas que tienen los atributos relevantes. Por ejemplo, encontrar una localización apropiada donde un centro comercial o un parque de diversiones puede ser establecido.
- Consultar los atributos geográficos de una localización específica. Por ejemplo, examinar los caminos de una localidad en particular, verificar la densidad de la carretera o encontrar la ruta más corta.

Por otra parte un GIS debe tener la capacidad de dar respuesta a preguntas como las siguientes:

1. ¿Dónde está el objeto A?
2. ¿Dónde está A con relación al objeto B?
3. ¿Cuantas ocurrencias del tipo A hay en una distancia D de B?
4. ¿Cuál es el valor que toma la función Z en la posición *x*?
5. ¿Cuál es la dimensión de B (Frecuencia, perímetro, área, volumen)?
6. ¿Cuál es el resultado de la intersección de diferentes tipos de información?
7. ¿Cuál es el camino más corto (menor resistencia o menor costo) sobre el terreno desde un punto (*x1, y1*) a lo largo de un camino P hasta un punto (*x2, y2*)?

8. ¿Qué hay en el punto (*x, y*)?
9. ¿Qué objetos están próximos a aquellos objetos que tienen una combinación de características?
10. ¿Cuál es el resultado de clasificar los siguientes conjuntos de información espacial?

1.1.1 Formato para la Representación de la Información Espacial

La información geográfica con la cual se trabaja en los GIS puede encontrarse en dos tipos de formatos: Celular o Raster y Vectorial como se muestra en la figura 1.

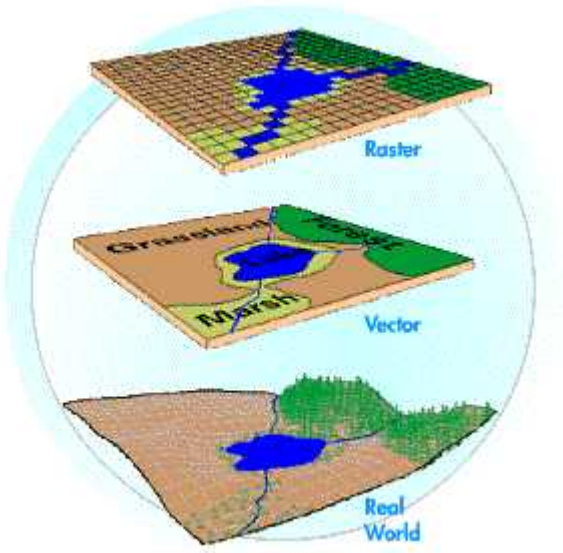


Figura 1.1 Formatos para representar la información espacial

Formato RASTER

El formato raster se obtiene cuando se "digitaliza" un mapa o una fotografía o cuando se obtienen imágenes digitales capturadas por satélites. En ambos casos se obtiene un archivo digital de esa información. Los archivos de imágenes se componen de celdas grid conocidas como píxeles [2]. La captura de la información en este formato se hace mediante los siguientes medios: scanners, imágenes de satélite, fotografía aérea, cámaras de video entre otros.

Formato VECTORIAL

La información gráfica en este tipo de formatos se representa internamente por medio de segmentos orientados de rectas o vectores. De este modo un mapa queda reducido a una serie de pares ordenados de coordenadas, utilizados para representar puntos, líneas y polígonos. La captura de la información en el formato vectorial se hace por medio de: mesas digitalizadoras, convertidores de formato raster a formato vectorial, sistemas de geoposicionamiento global (GPS), entrada de datos alfanumérica, entre otros. Cuando se hace uso de este formato, el mundo está representado como un mosaico de líneas y puntos interconectados que representan la localización y los límites de entidades geográficas [2].

1.2 Datos Espaciales

Un modelo de datos geográfico es una abstracción del mundo real que emplea un conjunto de objetos, para dar soporte al despliegue de mapas, consultas, edición y análisis. Los datos geográficos, muestran la información relativa a mapas, que representan la geografía como formas geométricas, redes, superficies, ubicaciones e imágenes, a los cuales se les asignan sus respectivos atributos que los definen y describen.

Un dato espacial es una variable asociada a una localización del espacio. Normalmente se utilizan datos vectoriales, los cuales pueden ser expresados mediante tres tipos de objetos espaciales:

1. Puntos. Se encuentran determinados por las coordenadas terrestres medidas por latitud y longitud. Por ejemplo, ciudades, accidentes geográficos puntuales, postes.

2. Líneas. Objetos abiertos que cubren una distancia dada y comunican varios puntos o nodos, aunque debido a la forma esférica de la tierra también se le consideran como arcos. Líneas telefónicas, carreteras y vías de trenes son ejemplos de líneas geográficas.

3. Polígonos. Figuras planas conectadas por distintas líneas u objetos cerrados que cubren un área determinada, como por ejemplo países, regiones o lagos.

De esta forma la información sobre puntos, líneas y polígonos se almacena como una colección de coordenadas (x, y). La ubicación de una característica de tipo punto, puede describirse con un sólo punto (x, y). Las características lineales, pueden almacenarse como un conjunto de puntos de coordenadas (x, y). Las características poligonales, pueden almacenarse como un circuito cerrado de coordenadas.

La otra forma de representar datos espaciales es mediante rasterización, la cual, a través de una malla que permite asociar datos a una imagen; es decir, se pueden relacionar paquetes de información a los píxeles de una imagen digitalizada. Los datos espaciales además se caracterizan por su naturaleza georreferenciada y multidireccional. La primera se refiere a que la posición relativa o absoluta de cualquier elemento sobre el espacio contiene información valiosa, pues la localización debe considerarse explícitamente en cualquier análisis. Por multidireccional se entiende que existen relaciones complejas no lineales, es decir, que un elemento cualquiera se relaciona con su vecino y además con regiones lejanas, por lo que la relación entre todos los elementos no es unidireccional. Aunque todos los elementos se relacionan entre sí, existe una relación más profunda entre los elementos más cercanos.

1.2.1 Características de los Datos Espaciales

Los datos espaciales son considerados como una clase especial de datos, sus características principales son las siguientes [6]:

a) Los objetos espaciales tienen una estructura compleja.
Un punto, o un conjunto de varios polígonos pueden caracterizar un objeto espacial de dato. Las tuplas en las bases de datos relacionales con tamaño fijo no son convenientes

para almacenar tal variedad de formatos de datos. Como resultado, operaciones espaciales (por ejemplo: intersección o unión) son computacionalmente más caras que las operaciones estándar de un RDBMS.

b) Los datos espaciales son a menudo dinámicos.
Esta característica de los datos espaciales requiere estructuras de datos robustas para inserciones frecuentes, borrado y actualización de objetos.

c) Las bases de datos espaciales tienden a ser grandes.
El número de objetos en un mapa geográfico a menudo demanda muchos gigabytes de almacenamiento. La integración de una memoria secundaria en estructuras de datos espaciales es por consiguiente requerida.

d) No hay un álgebra espacial estándar
No tiene bien definido un conjunto de operadores espaciales estándar, estos usualmente dependen del dominio de la aplicación de la base de datos espacial específica.

e) Los operadores espaciales son no cerrados.
La intersección de dos objetos espaciales, por ejemplo, puede regresar un conjunto de puntos, líneas o regiones.

Otra característica importante concerniente a datos espaciales, es que son multidimensionales, esto hace muy difícil aplicar métodos tradicionales de indexado.

1.2.2 Calidad de los Datos Espaciales

La calidad de los datos espaciales esta en función de su origen y captura. Son referentes de la calidad de los datos geográficos [7]:

1. Exactitud posicional. Diferencia entre la ubicación real y la del mapa. A nivel de representación es usual que el error máximo posicional no deba superar el 1/2 mm a la escala de impresión o salida final.

2. Consistencia y exactitud temáticas. La consistencia temática se refiere a la completa y correcta asignación de los atributos no espaciales a los objetos. Un error de ejemplo sería que se digitalizase un lago pero se señala en la base de datos como otro objeto o con características diferentes. Otro error es dejar accidentalmente objetos sin identificar. La exactitud temática se refiere a que tan aproximado es el valor del atributo considerado. Por ejemplo, algunas ciudades estarán identificadas en la base de datos con una población específica pero en la interfaz gráfica pueden representarse en rangos de 500 habitantes. Mientras que en la base de datos la aproximación es a la unidad, en la representación puede ser igual o mayor.

3. Consistencia topológica. Se relaciona con los errores de tipo topológico, por ejemplo: que los arcos se intersecten efectivamente en nodos cuando se desea representar intersecciones (de lo contrario simularán puentes). Evitar la doble digitalización y mucho menos si no es precisa.

4. Temporalidad -Historia de los datos-. Se refiere a un conocimiento de la actualidad de la información: conocer los tiempos en que cada fuente reporta los datos.

5. Integridad de los datos -Exhaustividad-. Comprende todo el proceso de estructuración de la información a partir de un correcto modelamiento y planificación de los datos a capturar: recolección de los datos necesarios, correcta asignación de relaciones entre entidades, escalas de trabajo adecuadas, etc.

Otros errores corresponden al proceso de empalme (Edge matching) donde pueden haber capas mal empatadas espacialmente y/o que coincidan figuras de distinta identidad, como lo muestra la figura 1.2, en un ejemplo de Arcview al hacer empalme.

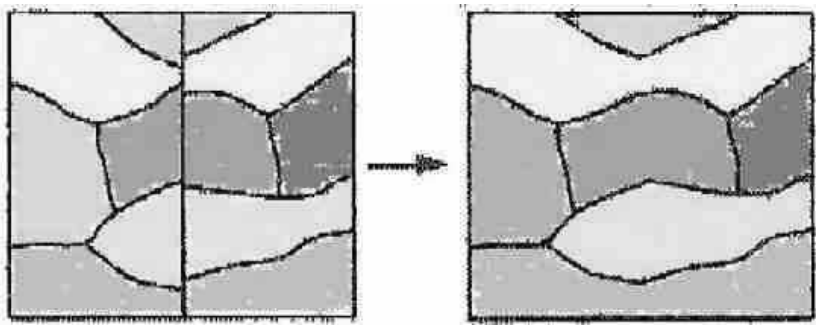


Figura 1.2 Empalme en ArcView de dos capas

En la figura 1.2 se muestra un ejemplo de datos originales distorsionados que propagan el error en su captura. Esto sucede cuando se digitalizan mapas arrugados o se trabaja con fotos aéreas no corregidas.

1.3 Regiones Vagas

Las regiones vagas son aquellas cuyos límites no se encuentran o no pueden ser precisamente definidos [2]. La representación de este tipo de regiones ayuda a tener una mejor definición de alguna zona para la cual no se puede definir con un grado de certidumbre cierto tipo de pertenencia a una clase de objetos.

Cuando se modelan datos espaciales y se asume implícitamente que la extensión y el límite de los objetos espaciales está determinado de forma precisa se habla de modelos objetos exactos. Los puntos, líneas y polígonos son representados en un marco de referencia dado, las líneas enlazan una serie de puntos (coordenadas) exactamente conocidas, y las regiones están limitadas por líneas exactamente bien definidas los cuales son denominados límites. Las propiedades del espacio de los puntos, a lo largo de las líneas, o dentro de las regiones son dadas por atributos cuyos valores se asumen como constantes sobre la magnitud total de los objetos [2]. Algunos ejemplos de objetos espaciales determinados de una forma precisa son: carreteras, caminos, casas y puentes [8].

Por otro lado existen los objetos espaciales indeterminados, cuya característica está dada por el esclarecimiento del límite o borde del objeto, el cual no está bien definido. Por esto existen las siguientes dos categorías de límites indeterminados [8]:

a) Límites cuya posición y forma son desconocidos o no pueden ser medidos con precisión.
b) Límites que no están bien definidos (por ejemplo, entre una montaña y un valle) donde esencialmente las relaciones topológicas entre objetos espaciales son de interés.

De acuerdo a esta clasificación, dos clases de vaguedad (fuzziness) o indeterminación concernientes a objetos espaciales se distinguen:

a) Vaguedad Posicional: se refiere a la incertidumbre relacionada con la carencia de conocimiento acerca de la posición y forma de un objeto con un borde real existente.

b) Vaguedad de dimensión: Se refiere a la incapacidad de medir un objeto de forma precisa.

Lo difuso en una región se refiere al hecho de que un objeto no tiene un borde bien definido, a pesar de que el objeto cuente con límites los cuales no han sido establecidos de forma precisa. Esta característica describe la vaguedad de un objeto, la vaguedad espacial ha sido tratada por geógrafos pero más cuidadosamente por científicos de la computación [2]. Por lo menos tres métodos han sido propuestos [8]:

Modelos exactos: transfieren los modelos de datos, tipos de sistemas y conceptos para objetos espaciales con límites a objetos espaciales sin límites claros y predominan modelos de incertidumbre pero también aspectos difusos. Un beneficio de los modelos exactos es que existen definiciones, técnicas, estructuras de datos, algoritmos, etc., que no necesitan ser desarrolladas sino solamente modificadas y extendidas, o simplemente usadas [8].

Modelos probabilísticos: están basados en la teoría de la probabilidad, predominantemente el modelo posicional y medidas de incertidumbre. Esta teoría es utilizada para representar incertidumbre. Define el grado de membresía de una entidad a un conjunto a través de una función de probabilidad estadísticamente definida [8].

Modelos difusos: todo se basa en la teoría de conjuntos difusa y predominantemente modelos difusos [8].

Los conjuntos difusos fueron introducidos por primera vez en 1965 por Lotfi A. Zadeh, quien es denominado por eso el padre de toda esta teoría [9]. La lógica difusa provee por sí misma un medio para acoplar a tareas como los conjuntos difusos. En cierta forma, la lógica difusa puede ser vista como un lenguaje que permite trasladar oraciones del lenguaje natural a un lenguaje matemático formal. Al principio el objetivo de esta área era ayudar a manejar aspectos imprecisos del mundo real, el descubrimiento de la lógica difusa permitió el desarrollo de aplicaciones prácticas. La teoría difusa es una extensión o generalización de la teoría booleana. Los conjuntos difusos permiten obtener el grado de pertenencia de un miembro al conjunto; ejemplos de objetos espaciales difusos incluyen montañas, valles, océanos y muchas otras características geográficas las cuales no pueden ser definidas de manera exacta por algún objeto espacial o forma.

1.3.1 Modelos Difusos

Dos tipos de lógica son usados y aplicados en el procesamiento de datos espaciales: la lógica booleana y la lógica difusa [10]. Después de que el profesor Zadeh en 1965 usará los conjuntos difusos, le siguieron Robinson y Strahler (1984) y Burrough (1986) para impulsar el uso de los conjuntos difusos en el manejo de la incertidumbre con bases de datos geográficas [2].

Existen dos formas de incertidumbre [10]:

a) Visibilidad probable, reconoce la incertidumbre si una localización es visible (esto es hay una línea directa de visión entre observar y localizar).
b) Claridad Difusa, define el grado en que algún objetivo potencialmente visible pueda ser distinguido.

La lógica difusa modela el espacio humano de una manera más cercana a la realidad. Las personas no perciben su ambiente como teniendo límites fijos, ya que se pueden mover de un lugar a otro en cuestión de un tiempo t. Por ejemplo, una persona se puede mover de una zona A a una zona B en cuestión de 10 minutos, 10 minutos es un fenómeno booleano sin duda, pero en el uso coloquial de un idioma "10 minutos" pocas veces significa eso, probablemente signifique menos de 20 minutos y más de 5 minutos, este concepto es actualmente mucho más cierto, y también es difuso, la idea de la proximidad también es difusa [2].

1.3.2 Lógica Difusa

La palabra fuzzy viene del inglés fuzz (tamo, pelusa, vello) y se traduce al español por difuso o borroso [9]. En la actualidad es un campo de investigación muy importante, tanto por sus implicaciones matemáticas o teóricas como por sus aplicaciones prácticas. Muchos conceptos que manejamos los humanos a menudo, no tienen una definición clara, por ejemplo podemos hacer este tipo de preguntas ¿Qué es una persona alta?, ¿A partir de qué edad una persona deja de ser joven?, y tener un rango de respuestas para estas dos preguntas simples.

Por otra parte la lógica booleana o bivaluada suele ser demasiado restrictiva, ya que una afirmación puede no ser ni cierta ni falsa [9], por ejemplo en la oración "yo leeré el Quijote" ¿en qué medida es cierto?, depende de quién lo diga y de que en "verdad" lo haga. En un amplio sentido, la lógica difusa se refiere a todas las teorías y tecnologías que emplean conjuntos difusos, que son clases con límites abruptos [11], [12].

Conceptos Imprecisos

Cuando se hace una oración para realizar una afirmación, existen conceptos en ella que no son muy claros para un lenguaje matemático, pero sin embargo es la forma de nuestro lenguaje coloquial y nuestra percepción del mundo real, a continuación se muestran ejemplos cotidianos del habla hispana:

La temperatura está caliente
La inflación actual aumenta rápidamente
Los grandes proyectos generalmente tardan mucho
Nuestros precios están por abajo de los precios de la competencia

Microsoft es una compañía grande y agresiva
Alejandro es alto pero Ana no es bajita

Estas proposiciones forman el núcleo de nuestras relaciones con "la forma de ver las cosas en el mundo". Sin embargo, son incompatibles con el modelado tradicional y el diseño de sistemas de información. Si podemos incorporar estos conceptos logramos que los sistemas sean potentes y se aproximen más a la realidad.

Por último, la tecnología difusa se debe o se recomienda utilizar [9], en los siguientes casos:

– En procesos complejos, si no existe un modelo de solución sencillo.
– En procesos no lineales.
– Cuando haya que introducir la experiencia de un operador "experto" que se base en conceptos imprecisos obtenidos de su experiencia.
– Cuando ciertas partes del sistema a controlar son desconocidas y no pueden medirse de forma fiable (con errores posibles).
– Cuando el ajuste de una variable puede producir el desajuste de otras.
– En general, cuando se quieran representar y operar con conceptos que tengan imprecisión o incertidumbre (como en las Bases de Datos Difusas).

1.3.3 Conjuntos Difusos

Los conjuntos difusos surgieron como una nueva forma de representar la imprecisión y la incertidumbre, un conjunto difuso es aquel cuyos límites no pueden ser definidos de una forma precisa. La teoría de conjuntos difusos es una generalización de la teoría de conjuntos clásica para permitir membresías parciales. Los conjuntos clásicos tienden a ser restrictivos y su objetivo es el de clasificar objetos y conceptos, por ejemplo, el conjunto de colores primarios: {Azul, Amarillo, Rojo}, pero existen ciertos conjuntos que no tienen un límite claro, por ejemplo, decimos que alguien es alto si tiene una estatura de 1.80 m, sin embargo alguien que mida 1.79 m ¿pertenecerá al conjunto de los bajos?, entonces los conjuntos difusos ayudan a suavizar la pertenencia a un conjunto y no obedecer la regla de estar o no estar en un conjunto (como sucede en los clásicos).

La teoría de los conjuntos difusos direcciona esta limitación permitiendo una membresía que admita obtener el grado de pertenencia de un objeto al conjunto. El grado de membresía en un conjunto se expresa por un número entre 0 y 1; donde 0 significa que no se encuentra en el conjunto, números entre 0 y 1 significan que parcialmente se encuentran en el conjunto y 1 significa que completamente está en el conjunto. Con ello se permite una transición gradual desde las regiones fuera del conjunto a aquéllas en donde el conjunto puede ser descrito.

Un conjunto difuso está definido por una función que mapea objetos en un dominio de acuerdo a su valor de pertenencia en el conjunto [11], [12]. Esta función es llamada función de pertenencia o membresía.

Función de Pertenencia o Membresía

Un conjunto difuso puede representarse también gráficamente como una función, especialmente cuando el universo de discurso X (o dominio subyacente) es continuo (no discreto).

– Abcisas (eje X): Universo de discurso X.
– Ordenadas (eje Y): Grados de pertenencia en el intervalo [0,1].

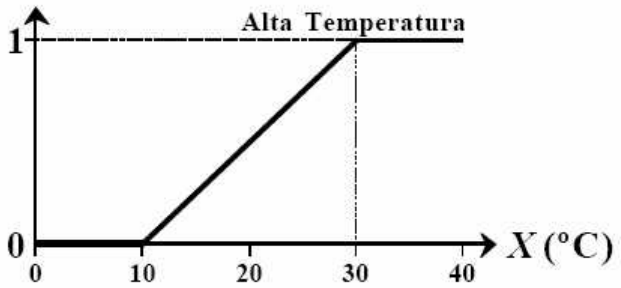


Figura 1.3 Concepto de temperatura Alta mapeada en una función de membresía [9].

En la figura 1.3 se representa el concepto de "Temperatura Alta", cuando en el eje X se alcanza un valor igual a 30, al eje Y le corresponde un valor de 1, lo cual significa que a partir de 30 ºC se habla de temperatura alta.

Una forma de expresar una función de membresía, de manera formal será:

$$F(x) = \begin{cases} 0 & \text{si } \notin \text{ a la región} \\ [0,1] & \text{si se encuentra en el borde de la región} \\ 1 & \text{si } \in \text{ a la región} \end{cases}$$

Existen diferentes tipos de funciones de membresía las cuales son :

Triangular
Γ (gamma)
S
Gausiana
Trapezoidal
Pseudo-Exponencial
Trapecio Extendido

La más comúnmente utilizada es la trapezoidal por adaptarse de forma natural a la mayoría de los problemas.

1.3.4 Definición de una Región Vaga

Una Región Vaga como ya se ha mencionado, es aquella cuyos límites no están o no pueden ser definidos. La indeterminación de estas regiones se asocia a cambios temporales (frío, calor, lluvia, etc.) y está formada por un par de regiones disjuntas que son el núcleo y el límite o borde [3]. El núcleo, describe la parte predominante de la región vaga, es decir, el área que definitivamente pertenece y siempre va a pertenecer a ella. El límite o borde, describe la parte vaga, esto es, el área que no se puede saber con certeza si pertenece a la región o no.

Para poder aproximar y definir la parte que le corresponde al núcleo y la que corresponde al límite, se hace uso del Rectángulo de Mínimo Acotamiento (MBR). Para poder asignar los grados de pertenencia de una característica geográfica dentro de la región vaga en [13] se hace uso de una representación difusa llamada Rectángulo de Mínimo Acotamiento Difuso (FMBR).

1.3.5 Rectángulo de Mínimo Acotamiento

Una Región Vaga se define cuando sus bordes no están precisamente definidos, para la representación de este tipo de regiones, se debe de separar por sus dos componentes principales: el núcleo y el límite. El núcleo y el límite son aproximados por su rectángulo de mínimo acotamiento (MBR) respectivamente.

Un Rectángulo de Mínimo Acotamiento (MBR) se define como el rectángulo más pequeño orientado a los ejes X e Y, que encierra o bordea un objeto (por ejemplo: característica geográfica, conjunto de datos geográficos), esta especificado por dos coordenadas: xmin, ymin y xmax, ymax. El uso de MBRs en bases de datos geográficas es ampliamente practicado como un camino eficiente para poder localizar y acceder objetos en el espacio [2]. Otra ventaja en el uso del MBR es que todos los objetos pueden ser distinguidos con el mismo nivel de dimensionalidad, es decir, las características de puntos, líneas y áreas son todos representados como objetos 2-D [13]. Un ejemplo de un MBR es mostrado en la figura 1.4.

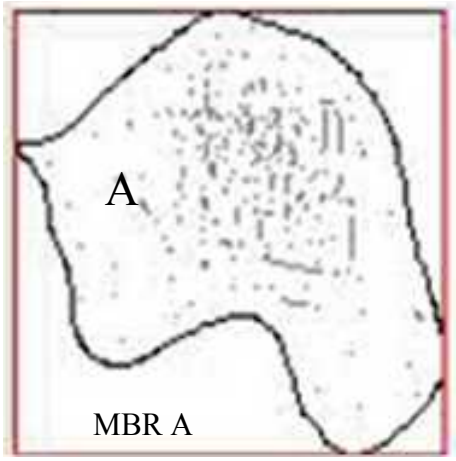


Figura 1.4 MBR de A

1.3.6 Rectángulo de Mínimo Acotamiento Difuso (Fuzzy MBR)

La representación difusa, llamada Rectángulo de Mínimo Acotamiento Difuso (FMBR), ha sido desarrollada para representar los diferentes grados de membresía de un punto localizado dentro de la región vaga [13].

El FMBR encierra todos los puntos del mapa de espacios donde una característica de interés es encontrada. Una característica se define como una entidad con atributos y relaciones comunes. El FMBR puede ser considerado también como el rectángulo circunscrito (CR) de un polígono difuso. La generación de rectángulos internos es realizada iterativamente hasta que se obtiene el rectángulo inscrito (IR) del objeto base. Así, el IR es el máximo rectángulo interno dentro del objeto y corresponde al centro de

la región difusa. La distancia entre el IR y el FMBR es utilizada para representar el límite difuso [13].

Una función de membresía espacial basada en la distancia Euclideana puede ser usada para determinar el grado de pertenencia de una característica al conjunto difuso. De esta manera, características dentro del IR tendrán un grado de membresía de 1. Este grado será gradualmente decrementado mientras más lejos se esté del centro. Los puntos localizados fuera del FMBR tendrán un grado de membresía de 0 [13].

La representación gráfica de un FMBR, es mostrada en la figura 1.5. La base de la región vaga A es aproximada por el FMBR (A). Esta primera aproximación se llama rectángulo circunscrito (CR) de una región difusa. El FMBR o CR corresponde al rectángulo mínimo con bordes paralelos al eje X e Y que encierran de forma óptima a la región vaga A [13].

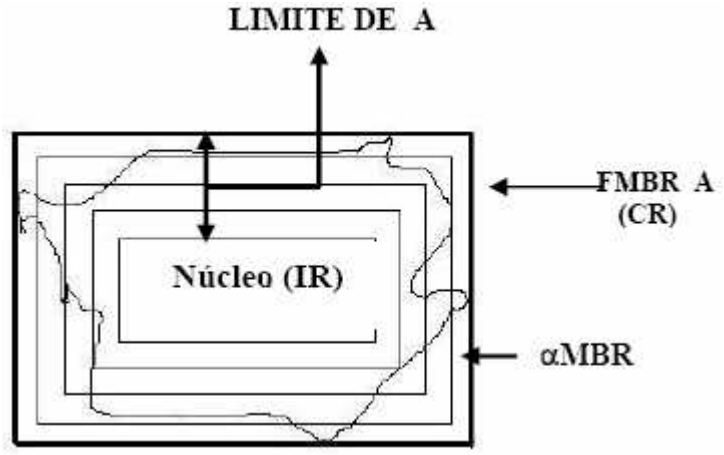


Figura 1.5 Representación de un FMBR

Un corte-αMBR (denominado α-cut) permite hacer finas distinciones dentro de la región vaga, los cortes-αMBR son regiones individuales dentro del FMBR, de esta manera, se puede pensar en una región estructurada difusa como una agregación de regiones α-nivel. Los αMBRs empiezan a definirse desde el borde del FMBR(A) al centro de A. El más externo de los corte-αMBR tiene el más bajo de los grados de membresía, el grado va en aumento conforme se acerca al núcleo cuya membresía es 1 [4], [12], [13].

En este trabajo de investigación se tiene contemplado utilizar el FMBR para representar los fenómenos geográficos asociados a la representación del borde, ya que el FMBR es adecuado para realizar dicha tarea sobre este tipo de regiones complejas, porque el MBR encierra de manera más aproximada las regiones vagas y para su representación sólo se necesitan conocer dos de los puntos del MBR (Xmin,Ymin y Xmax,Ymax) como se menciona en [13], y mediante el uso de funciones de membresía asignar los grados de pertenencia que ayudan a definir la representación del borde.

1.4 Almacén de Datos (Data Warehouse -DW)

De manera general se define a un Almacén de Datos como una colección de datos orientada a un dominio que tiene las características de ser temático (se enfoca a un tema en especial), integrado (toda la información que extrae de fuentes externas e internas es

integrada en una sola), variable en el tiempo y no volátil (se refieren a que los datos no serán actualizados ni borrados del almacén sino que se insertaran nuevos valores que tendrán una referencia con respecto a un periodo de tiempo) [14] y ayuda o brinda soporte en el proceso de toma de decisiones (usado mucho por las empresas u organizaciones). Por ejemplo, en el caso de las empresas, los almacenes sirven como expedientes diseñados para favorecer el análisis y la divulgación de manera eficiente de los datos. El almacenamiento de los datos no debe usarse con datos de uso actual. Los almacenes de datos manejan e integran grandes volúmenes de información [14] para facilitar los procesamientos analíticos posteriores.

Periódicamente, se importan los datos de sistemas del planeamiento del recurso de la empresa (ERP) y de otros sistemas de software relacionados al negocio en el almacén de los datos para la transformación posterior. Se acostumbra normalizar los datos antes de combinarlos en el almacén de datos, esta fase se suele realizar con una herramienta extracción, transformación y carga (ETL).

En la realidad, la carga y mantenimiento de un almacén de datos es uno de los aspectos que más esfuerzos requiere. ETL produce un repositorio de datos intermedio, como se muestra en la figura. 1.6 que realiza las funciones de extracción de las fuentes de datos (transaccionales ó externas), transformación (limpieza y transformación de datos, consolidación, creación de claves y mantenimiento de los metadatos) y la carga del Almacén de Datos (Indización, planificación de la carga y mantenimiento, pruebas de calidad).

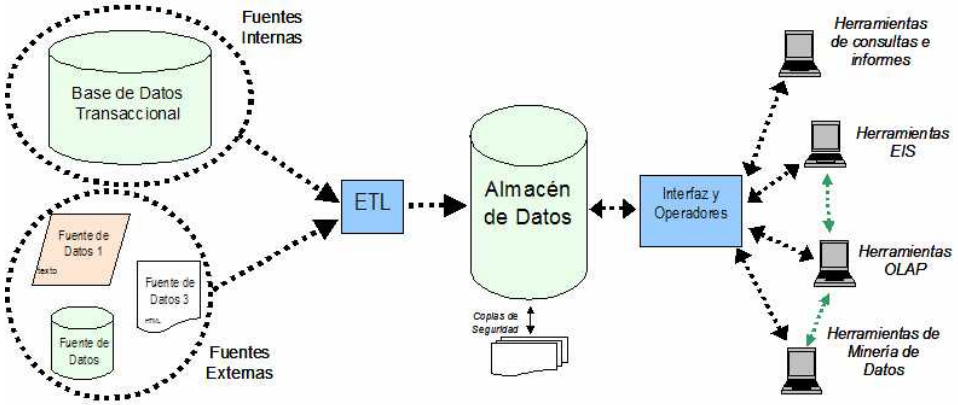


Figura 1.6. Perspectiva general de un Almacén de Datos.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas [15], como se muestra la figura 1.6.

Un Almacén de Datos debe entregar la información correcta a la gente indicada en el momento adecuado en el formato correcto. El Almacén de Datos da respuesta a las necesidades de los usuarios, utilizando Sistemas de ayuda en la toma de decisiones (DSS), Sistemas de información para ejecutivos (EIS) o herramientas para hacer consulta o informes. Los usuarios finales fácilmente pueden realizar consultas sobre sus Almacenes de Datos sin tocar o afectar la operación del sistema.

Un almacén de datos es una estructura ideal, para contener datos espaciales, pues se habla de una cantidad importante de información (grandes volúmenes de datos), aunado

al hecho de que la información almacenada contendrá relaciones a conjuntos difusos, lo cual incrementará aun más el tamaño de la información a procesar. El almacén de datos permitirá que toda esta información pueda ser consultada y representada en tiempos de respuesta satisfactorios al usuario

1.4.1 DataMarts

Un DataMart es una vista lógica de datos sin procesar provistos por el sistema de operaciones hacia el Datawarehouse con la adición de nuevas dimensiones o información calculada. Se les llama DataMart, porque representan un conjunto de datos relacionados con un tema en particular como Ventas, Operaciones, Recursos Humanos, etc., y están a disposición de los "clientes" a quienes les pueden interesar. Esta información puede accesarse por el Administrador mediante "Tablas Dinámicas" ó programas personalizados.

Las Tablas Dinámicas le permiten manipular las vistas (cruces, filtrados, organización) de la información con mucha facilidad. Los cubos de información (DataMarts) se producen con mucha rapidez. La información estratégica está clasificada en: Dimensiones y Variables. El análisis está basado en las dimensiones y por lo tanto es llamado: Análisis multidimensional. Llevando estos conceptos a un DW: Un Data Warehouse es una colección de datos que está formada por Dimensiones y Variables, entendiendo como Dimensiones a aquellos elementos que participan en el análisis y Variables a los valores que se desean analizar.

Los datamarts son utilizados para transformar los almacenes de datos en subdivisiones lógicas de bases de datos. El beneficio de esto, es que los almacenes de datos no se sobrecargan con las tareas de consultas. Únicamente se manejan datos relevantes, por lo tanto cuando las consultas son elaboradas toman menos tiempo que ser ejecutadas desde el mismo almacén de datos.

Muchos confunden esta estructura con los cubos de datos, ya que ambos son clasificados como modelos de datos multidimensionales y por que también ambos almacenan medidas y dimensiones, sin embargo, un datamart se especializa únicamente en un tema o cumple solo la tarea para la que está asignado. El cubo de datos por su parte hace un precálculo de todas las combinaciones de las medias alrededor de todas las dimensiones calculadas y también hace un precálculo de todas las agregaciones. Un cubo de datos puede ser alimentado directamente de un Almacén de Datos o cualquier otra fuente de datos, no es necesario generar un esquema de estrella o copo de nieve para alimentar al cubo. Un datamart con un esquema permitirá en un cubo de datos una recuperación de información eficiente, ya que la alimentación del cubo es más rápida, aunque ello indique que el cubo de datos será creado para un tema específico.

1.4.2 Elementos que conforman un Almacén de Datos

Metadata

Uno de los componentes más importantes de la arquitectura de un DW es el Metadata. Es definido comúnmente como "datos acerca de los datos", en el sentido de que se trata de datos que describen cuál es la estructura de los datos y cómo se relacionan. El Metadata documenta exactamente, entre otras cosas, qué tablas existen para esa aplicación, qué columnas posee cada una de las tablas y qué tipo de datos se pueden almacenar. Los datos son de interés para el usuario final, el Metadata es de interés para los programas que tienen que manejar estos datos. Sin embargo, el rol que cumple el Metadata en un ambiente de DW es muy diferente al rol que cumple en los ambientes operacionales. En un ambiente de DW el Metadata juega un papel fundamental, ya que también se llama diccionario en el ambiente operacional y es la base de la estructura del Almacén que se vaya a formar.

Middleware

La función del Middleware es la de asegurar la conectividad entre todos los componentes de la arquitectura de un DW. El Middleware puede verse como una capa API (Application Programmer Interface - Interfaz de Programación de Aplicación), en base a la cual los programadores pueden desarrollar aplicaciones que trabajen en diferentes ambientes sin preocuparse de los protocolos de red y comunicaciones en que se correrán. De esta manera se ofrece una mejor relación costo/rendimiento que pasa por el desarrollo de aplicaciones más complejas, en menos tiempo.

API

Lenguaje y formato de mensaje utilizados por un programa para activar e interactuar con las funciones de otro programa o de un equipo físico. Asegura la conectividad entre todos los componentes de una infraestructura informática. Es la estructura para enlazar todas las aplicaciones en forma integrada.

Mecanismos de Extracción

Otro de los componentes de la arquitectura de un DW son los sistemas OLAP. Estos tipos de sistemas están orientados a la realización de análisis estratégicos de la información contenida en un DW de una manera ad-hoc. Los análisis estratégicos requieren de una visión dinámica y multidimensional de la información diferente a la que se encuentra en los sistemas OLTP. Este tipo de análisis está orientado a procesar grandes volúmenes de datos para poder medir la evolución del negocio a través del tiempo, mediante la realización de comparaciones, el estudio de indicadores, desviaciones, etc. Esto requiere la posibilidad de realizar análisis Top Down, es decir que estos sistemas deben poseer el dinamismo necesario para permitir la reformulación de la consulta realizada de acuerdo al análisis de los resultados obtenidos en una primera instancia.

Mecanismos de Carga

Existen dos formas básicas de desarrollar esta tarea:

Acumulación Simple.- La acumulación simple es, sin duda, la más sencilla y común, y consiste en realizar una sumarización o resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el DW.

Rolling.- El proceso de Rolling por su parte, se aplica en los casos en que se opta por mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo.

Otra de las características que tienen los Almacenes de Datos es que son de sólo lectura, a diferencia de una base de datos convencional (OLTP), que realiza operaciones de inserción, actualización, borrado y lectura. El periodo de tiempo cubierto por un Almacén de Datos varía entre 2 y 10 años [15].

1.4.3 Arquitectura Multidimensional de un Almacén de Datos

Las herramientas que hacen uso de los almacenes de datos para llevar a cabo los análisis para lo que fueron creados, han adoptado un modelo multidimensional de datos, entonces el almacén es visto desde una perspectiva multidimensional como lo muestra la figura 1.7.

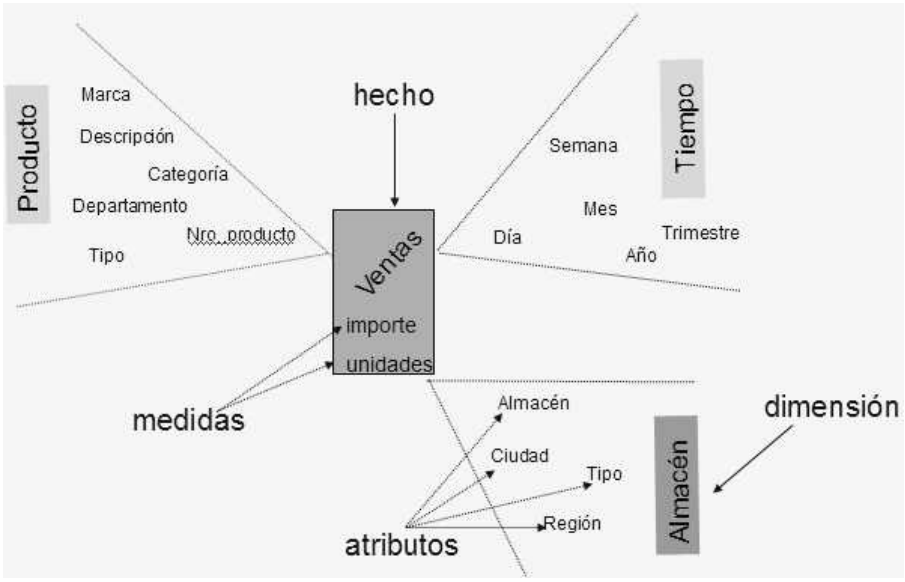


Figura 1.7 Arquitectura de un Almacén de Datos en un modelo multidimensional [15]

En este modelo, se presenta un tema (Ventas) que se quiere analizar, representado por la tabla de hechos y que va a proporcionar datos a través de medidas (unidades, importe de ventas), las medidas son obtenidas de las diferentes tablas denominadas dimensiones, las cuales van a estar relacionadas por la tabla de hechos mismas que se pueden organizar en una forma jerárquica. Dicha organización es por niveles, con lo cual se generan especializaciones hacia los datos que se estén analizando, a este proceso se le conoce como Agregación y Desagregación. Por ejemplo en la figura 1.7, la dimensión Tiempo se puede agregar por **Año**, después por **Trimestre**, luego seguiría **Mes**,

continuando por **Semana** y finalmente **Día**, en orden jerárquico descendente, al proceso contrario se le denominará por lo tanto Desagregación. Es decir, entre los atributos de una dimensión se definen jerarquías [15]. Este modelo multidimensional descrito se puede presentar en diferentes esquemas.

1.4.4 Esquema de Estrella (Star)

Un **esquema en estrella** es aquel que tiene una tabla de hechos que contiene los datos de análisis, y relaciona las tablas lookup o de dimensiones (ver figura 1.8). Este esquema es ideal por su simplicidad y velocidad para ser usado para análisis: DataMarts y EIS. Además, permite reducir el número de joins entre tablas y deja a los usuarios establecer jerarquías y niveles entre las dimensiones.

Finalmente, es la opción con mejor rendimiento y velocidad pues permite indexar las dimensiones de forma individualizada sin que repercuta en el rendimiento de la base de datos en su conjunto.

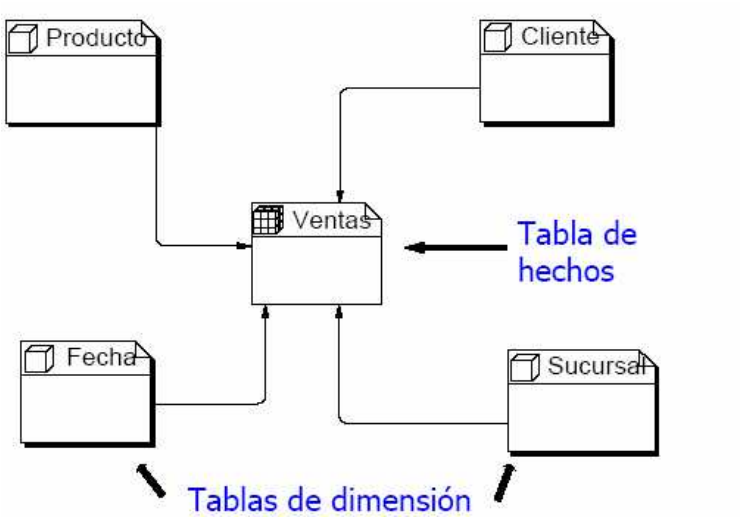


Figura 1.8 Esquema de Estrella de un Almacén de Ventas

1.4.5 Esquema de Copo de Nieve (SnowFlake)

Un esquema en copo de nieve es una estructura más compleja que el esquema en estrella. Se da cuando existen un gran número de tablas de hechos sin que sea factible reducir su número. Aunque puede reducir espacio, tiene la contradicción de peores rendimientos que el de estrella al tener que crear más tablas de dimensiones y más joins (relaciones entre las tablas) lo que tiene un impacto directo sobre el rendimiento.

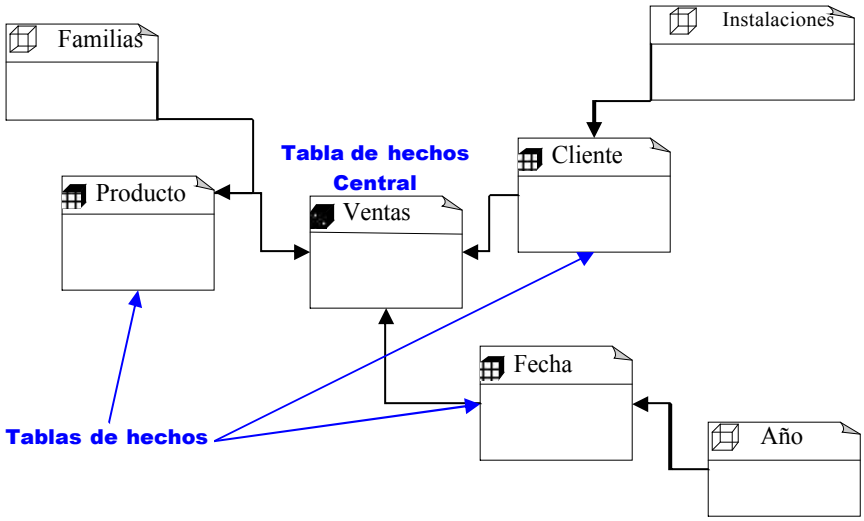


Figura 1.9 Esquema de Copo de Nieve

En la la figura 1.19, se representa un esquema Copo de Nieve, en donde Cliente, Provincias, Fecha y Producto son tablas de hechos que están relacionadas por la tabla de hechos central (Ventas), y las cuales a su vez relacionan las diferentes Dimensiones. Un esquema de copo de nieve puede ser visto como un árbol, donde su raíz es la tabla de hechos central o principal, los nodos de altura 1 son denominados dimensiones y los demás nodos de altura mayor a 1 son denominados subdimensiones. Si en esta estrucutura su altura mayor es igual a 1 entonces se denomina esquema de estrella [16].

1.4.6 Herramientas OLAP

Las herramientas OLAP presentan al usuario una visión multidimensional de los datos para cada actividad que es objeto de análisis. El usuario formula consultas a la herramienta OLAP seleccionando atributos de este esquema multidimensional sin conocer la estructura interna (esquema físico) del almacén de datos. La herramienta OLAP genera la correspondiente consulta y la envía al gestor de consultas del sistema (p.ej. mediante una sentencia SELECT). Una consulta a un almacén de datos consiste generalmente en la obtención de medidas sobre los hechos parametrizadas por atributos de las dimensiones y restringidas por condiciones impuestas sobre las dimensiones. En pocas palabras una herramienta OLAP permite obtener información generando consultas multidimensionales, con columnas y filas móviles y diversos grados de agrupamiento para diferentes parámetros (ejemplo de una herramienta OLAP: Mondrian).

Tipos de servidores OLAP.

MOLAP: Multimensional OLAP. Estos son arreglos multidimensionales que no escalan a grandes volúmenes. No hay estándar y es muy eficiente, cuya interfaz es estilo hoja de cálculo y se usa principalmente en operaciones de agregación de medidas diferentes, también tiene la capacidad de organizar los niveles jerárquicos de las dimensiones y realizar las operaciones de subir o bajar en los niveles de agregación (Roll-up, Drill-Down). Además realizan otras operaciones comunes: Filtrar y Rotar. (Slice and Dice).

ROLAP: Relational OLAP. Son elaborados a partir de SGBD Relacional y soportan consultas SQL, estos escalan bien a grandes volúmenes pero son menos eficientes. Los fabricantes de SGBD relacionales ofrecen extensiones y herramientas para poder utilizar el SGBDR como un Sistema Gestor de Almacenes de Datos.

HOLAP: Híbrido OLAP. En estos los datos agregados son soportados por el MOLAP y los datos detallados por el ROLAP.

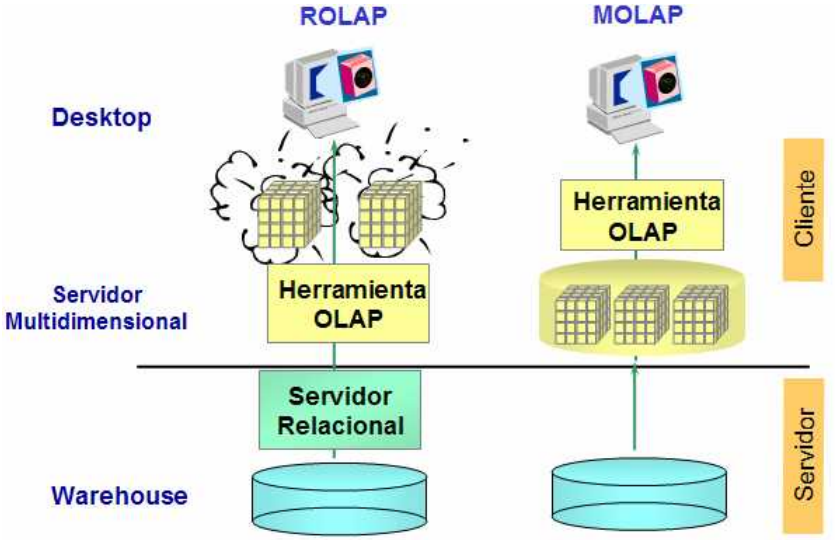


Figura 1.10 Tipos de OLAP: ROLAP y MOLAP.

En la figura 1.10 se observan las diferencias entre ROLAP y MOLAP. El ROLAP por su lado hace uso de un Servidor Relacional que funcione como Gestor del Almacén de Datos, luego haciendo uso de las Herramientas OLAP, optimizadores de consultas y extensiones SQL (operador CUBE, roll-up) generan la interfaz de salida al usuario final. MOLAP por su parte únicamente utiliza estructuras multidimensionales para que el servidor MOLAP presente dichas estructuras.

1.5 Estado del Arte

Han existido trabajos relacionados a la clasificación de los métodos de acceso espacial mediante el indexamiento [6], que han servido para la representación de datos espaciales en un almacén de datos espacial. En [23] se estudian los métodos para la elaboración de un análisis OLAP espacial, integrando los métodos no espaciales que maneja OLAP y la manipulación de bases de datos espaciales por medio de los métodos que se conocen para este propósito, también se propone un modelo de almacén de datos espacial así como los métodos para el cálculo de un cubo de datos espacial al cual se le puedan aplicar operaciones OLAP.

También ha existido la integración de lógica difusa dentro del análisis OLAP, como en [17], que parte de la generación de un cubo de datos difuso para la generación de reglas de asociación en el proceso de minería de datos, en [30] se presenta un modelo de un cubo de datos semántico haciendo uso de conjuntos difusos para la representación de la semántica lingüística de las dimensiones y de las medidas que componen al cubo de datos.

Todo este estudio ha servido para hacer uso de las diversas áreas que pueden abarcar las bases de datos espaciales y la lógica difusa en combinación con un almacén de datos, sin embargo, en la literatura no existe, al menos hasta lo investigado en este trabajo, un trabajo que trate de integrar ambas áreas, en el cuál se involucren datos espaciales con conjuntos difusos, dicha integración puede ser útil, este trabajo pretende describir el modelado en la generación y consulta de un cubo de datos espacial difuso a partir de la generación de un Almacén de Datos Espacial Difuso.

Existen trabajos previos en la Representación y Manipulación de Regiones Vagas. El más representativo es el FMBR [13], el cual incluye la utilización de Lógica Difusa para definir grados de pertenencia de acuerdo a la evaluación de una Función de Membresía integrando estas áreas con el uso del Rectángulo de Mínimo Acotamiento para encerrar la región a representar, ya que el MBR cuenta con demasiadas ventajas, las cuales ya se han mencionado en secciones anteriores.

Dentro de nuestro trabajo de investigación, también está la propuesta de representar un Cubo de Datos Difuso (CDD) el cual en [17] fue definido para un problema convencional (ventas). En este CDD, cada uno de los atributos de una base de datos relacional es representado mediante 2 o 3 funciones de membresía al cubo de datos, por lo tanto, es necesario extender este concepto para considerar la generación del CDD a partir de una Base de Datos Espacial.

El concepto del Cubo de Datos Espacial, ha sido usado en [18], para brindar soporte en el análisis multidimensional de datos relacionados a aspectos médicos. Los aspectos estadísticos y espaciales juegan un papel importante y son integrados para el desarrollo de dicho proyecto descrito.

CAPÍTULO 2

ANÁLISIS Y DISEÑO

Un Almacén de Datos como se menciona en [19] es un repositorio de datos orientado al tema, integrado, no volátil, y variable en el tiempo. Este concepto se ha extendido para integrar la geografía de los datos y conjuntos difusos, y así ayudar en el proceso de toma de decisiones y generación de conocimiento global.

En este capítulo abordaremos el análisis y diseño del almacén de datos espaciales difuso. Se partirá de la definición de la base de datos espacial de donde se extraerá la información que será cargada en el almacén y los conjuntos difusos necesarios que serán asociados a las características de las entidades en la base de datos espacial. Posteriormente se hará el análisis de las dimensiones que se requieran, se presentará el modelo multidimensional en un esquema de copo de nieve (snowflake), justificando la elección de este esquema para finalmente realizar el diseño del almacén de datos espacial difuso.

2.1 Diseño de un Almacén de Datos

Hoy en día existen varias propuestas en el diseño conceptual y lógico de los almacenes de datos, para representar las principales propiedades estructurales y dinámicas del modelado multidimensional. Sin embargo, ninguna propuesta ha sido aceptada como un modelo estándar, aunque se está trabajando en la generación de un modelo que pueda acercarse a un estándar o que englobe las características más importantes de las propuestas más interesantes en el diseño de un almacén de datos [20].

Para llevar a cabo nuestro diseño nos basamos en el propuesto en [15], el cual se compone de los pasos que se muestran en la figura 2.1.

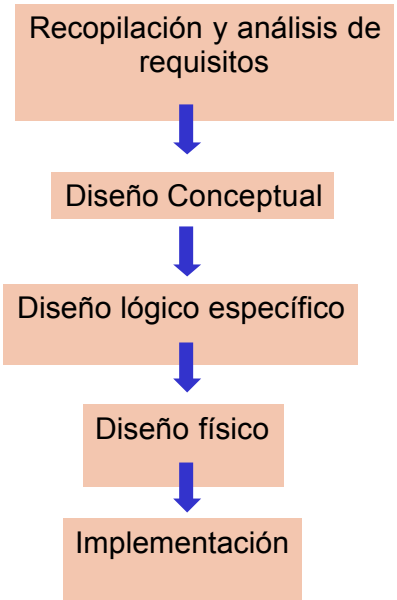


Figura 2.1 Etapas del Diseño de un Almacén de Datos en [15].

2.2 Recopilación y Análisis de Requisitos

Como ya se mencionó anteriormente, la información será recopilada de una base de datos espacial. Los conjuntos difusos que se generen a partir de ciertas funciones de membresía relacionadas a ciertos eventos de la base, asignarán grados de pertenencia a la característica que se quiera analizar. Toda la información que sea almacenada en la base de datos espacial, será proporcionada por los mapas del INEGI, a los cuales se les han aplicado cierto tipo de operaciones que nos han permitido filtrar la información requerida para la elaboración del almacén. Una vez obtenida la información a guardar, se analizarán las características vagas para representar los conjuntos difusos e integrar ambos conceptos dentro de las dimensiones de nuestro almacén en el modelo multidimensional.

El almacén de datos propuesto contendrá tres dimensiones:

- **Temporal**.- Esta dimensión contendrá los niveles de agregación Año, Semestre, Trimestre, Mes, Semana, Día que están relacionados a una fecha en que ocurre un evento.
- **Espacial**.- Esta dimensión contendrá la información georeferenciada, y se puede organizar de diferentes maneras, por ejemplo: Estado, Región, Municipio, Localidad, FMBR.
- **Temática**.- Estará enfocada a analizar las características geográficas de interés, por ejemplo, se puede medir el número de habitantes para un hecho: densidad demográfica (No. De Habitantes por kilómetro cuadrado). Las dimensiones que describen al número de habitantes podrán ser, tipo de suelo, vegetación, temperatura, hidrología y número de caminos.

2.3 Diseño Conceptual

Se explicará a continuación el modelo Entidad-Relación (ER) de la base de datos espacial denominada "Zonas de Riesgo en el Volcán Popocatépetl" y se presentarán algunos mapas de ArcView para extraer la información que será guardada en la base.

El modelo ER que se presenta en la figura 2.2, representa las entidades más importantes que van a participar en nuestro almacén de datos, este modelo se obtuvo de la información contenida en la página del CENAPRED, en la cual se hace referencia a mapas de peligros, las rutas de evacuación e información relacionada con el volcán, tomando la narrativa que viene en la página de Internet, se han extraído los elementos que se piensan, pueden intervenir en el análisis que en este trabajo se propone.

El modelo ER de la figura 2.2, servirá para diseñar el almacén de datos más adelante, tomando los elementos de las entidades que en el modelo participan, para formar una dimensión por cada una de ellas, teniendo en cuenta las características espaciales en aquellas entidades que contengan esta propiedad.

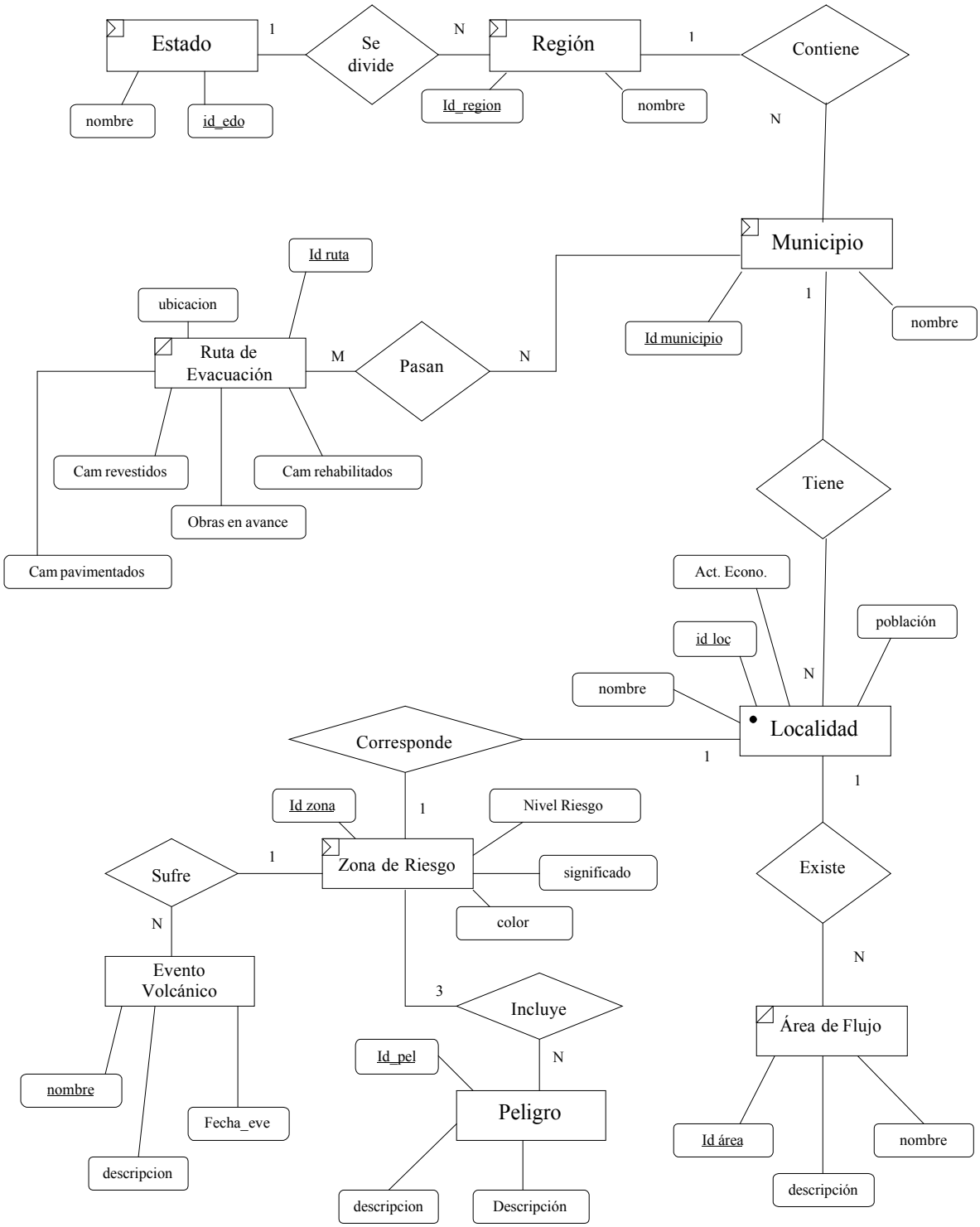


Figura 2.2 Modelo Entidad-Relación de la Base de Datos Espacial

La figura 2.2 muestra el modelo entidad relación de la base de datos propuesta para almacenar la información que está contenida en las capas de los mapas de Arcview. Cabe señalar que faltan atributos para las entidades pero por cuestiones de espacio no han sido representadas en este esquema. A continuación se presenta el mapa de peligros que ya ha sido clasificado por el INEGI, en lo que respecta a la zona del volcán Popocatépetl y zonas aledañas, que será información contenida en nuestra base de datos.

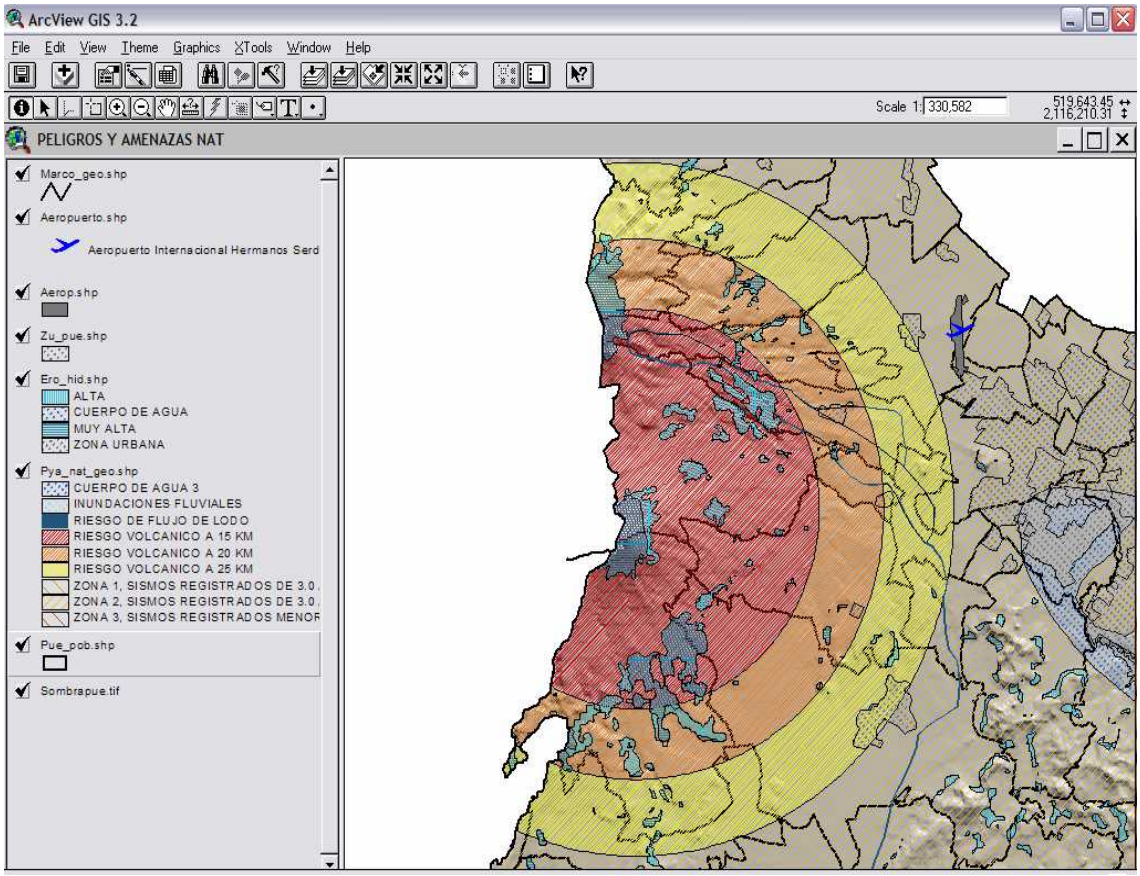


Figura 2.3 Zona de Riesgos del Estado de Puebla visualizada en ArcView

Como ya se mencionó, cuando se presenta un Almacén de Datos en un modelo multidimensional, este es descrito a partir de una tabla de hechos que representa el tema a analizar, la exploración se realiza mediante ciertos indicadores denominados medidas, las cuales son obtenidas de otras tablas que están relacionadas a la tabla de hechos y que en este contexto se conocen como dimensiones. Para más referencia ver la figura 1.7 del capítulo uno que presenta la arquitectura de un almacén de datos en un modelo multidimensional.

El modelado multidimensional se puede aplicar utilizando distintos modelos de datos (conceptuales o lógicos). La representación gráfica del esquema multidimensional dependerá del modelo de datos utilizado [15].

En la figura 2.4, se modela la arquitectura del almacén de datos tomando como referencia de la figura 1.7 del capítulo uno, es decir el almacén de datos acerca del Plan de Contingencia en el volcán Popocatépetl.

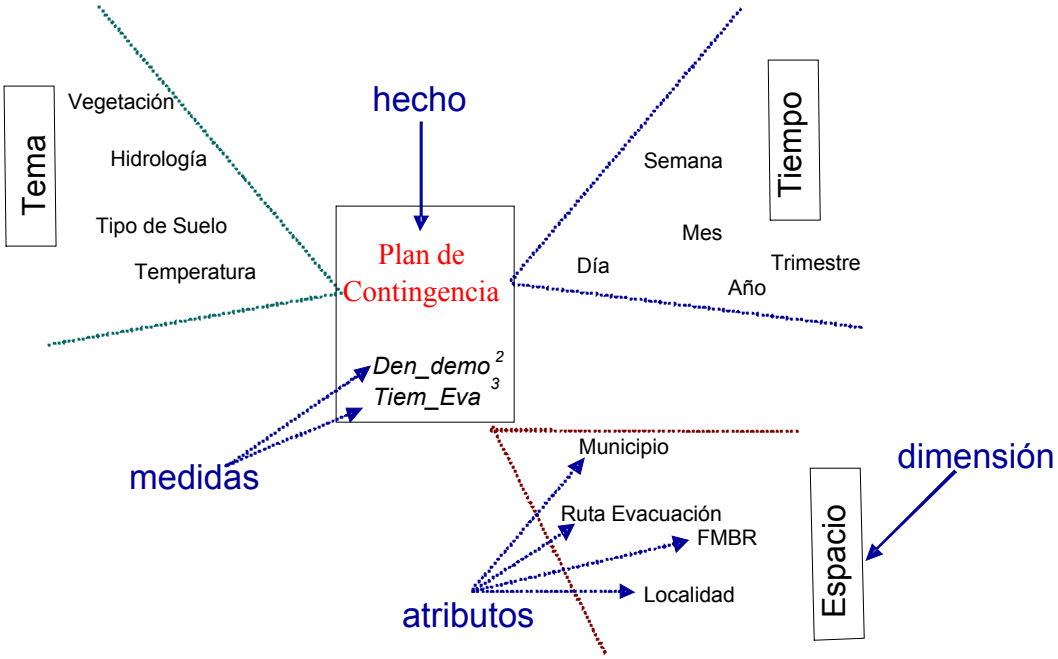


Figura 2.4 Arquitectura del Almacén de Datos Espacial Difuso

De la base de datos espacial se elijen los atributos que van a servir para generar la tabla de hechos para relacionar las dimensiones y medidas que estarán en el almacén. En la figura 2.5 se muestran los atributos clave de la tabla de hechos.

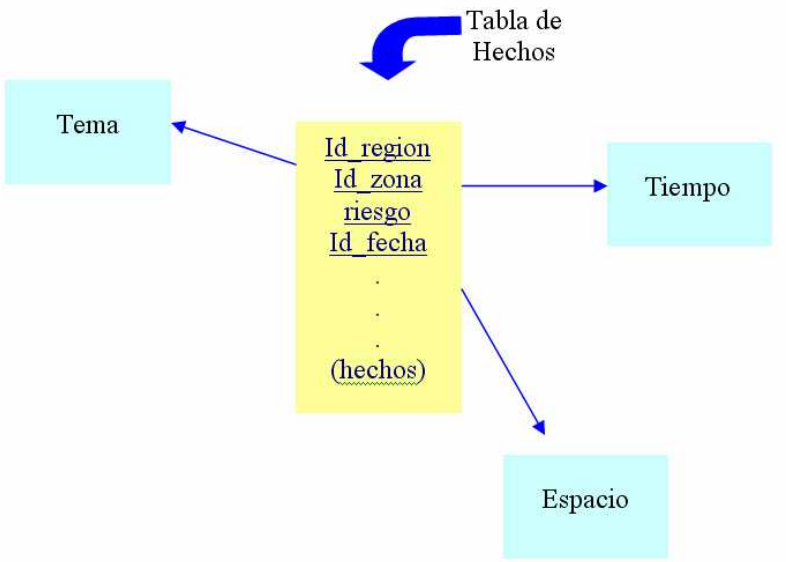


Figura 2.5 Elementos clave de la tabla de Hechos central

A su vez, Tema y Espacio se descomponen en otras tablas relacionadas por Tema y Espacio respectivamente, por lo que el esquema de copo de nieve es ideal para representar el modelo multidimensional del almacén de datos espaciales, en la figura 2.6 se representa el esquema de copo de nieve del almacén de datos espacial.

[1] Densidad Demográfica, [2] Tiempo de Evacuación

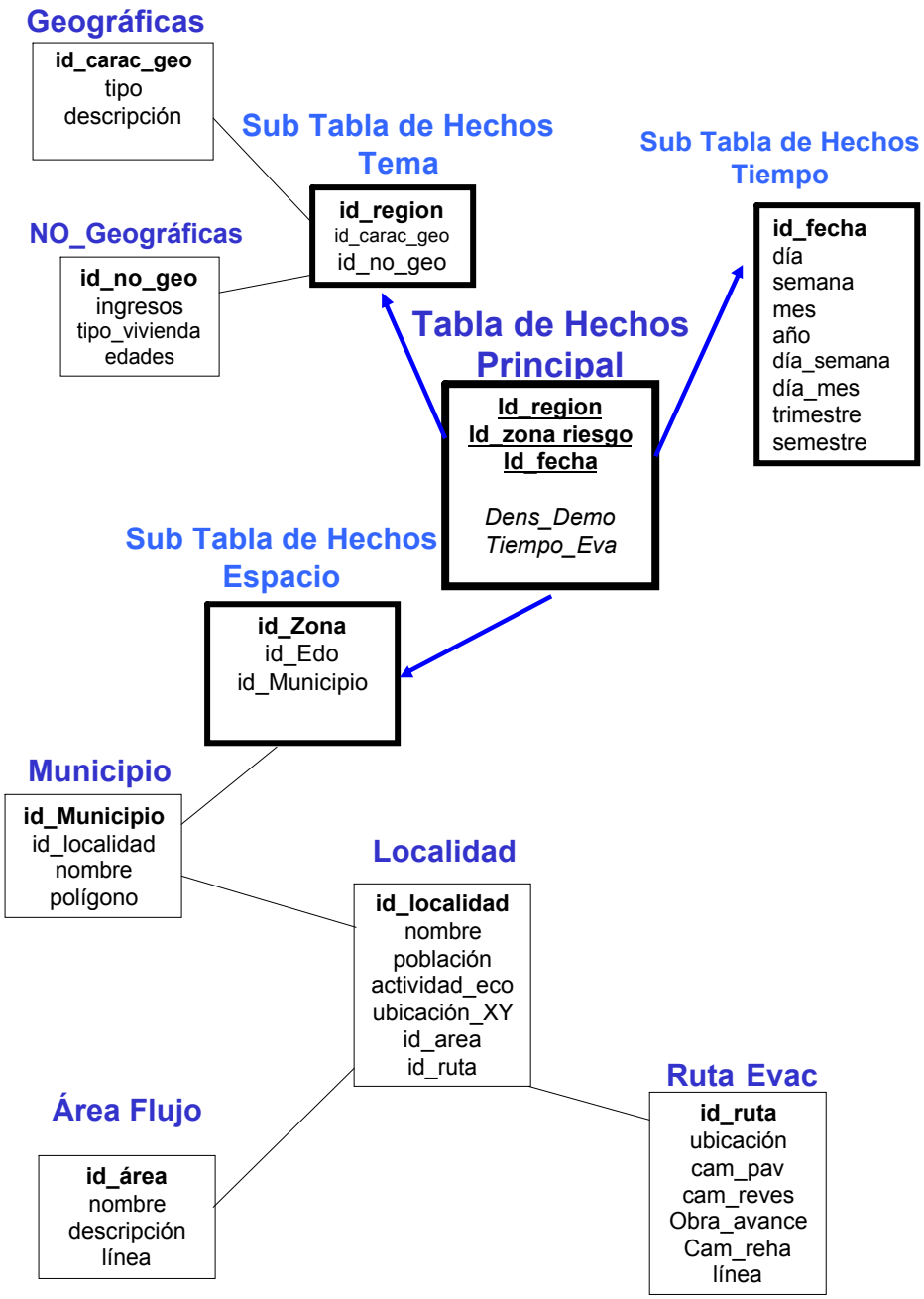


Figura 2.6 Esquema de Copo de Nieve del Almacén de Datos Espacial

En la figura 2.6 se representa el primer modelo de el esquema de copo de nieve propuesto el cual incluye elementos espaciales como lo son los atributos *polígono* en la tabla municipio de la dimensión Espacio y el atributo *ubicación_XY* de la tabla localidad de la misma dimensión.

La Tabla de Hechos Principal va a relacionar a las subtablas de hechos de cada dimensión, que a su vez relacionan a todas las tablas que están contenidas dentro de cada dimensión, de esta manera la información se va a integrar en el nivel 0, es decir en la tabla de hechos central o principal, a continuación se muestra el esquema a un nivel 1, es decir, con todas las subtablas de hechos y sus conjuntos difusos que se pretenden asociar a cada una de ellas mediante los atributos que necesiten ser más específicos.

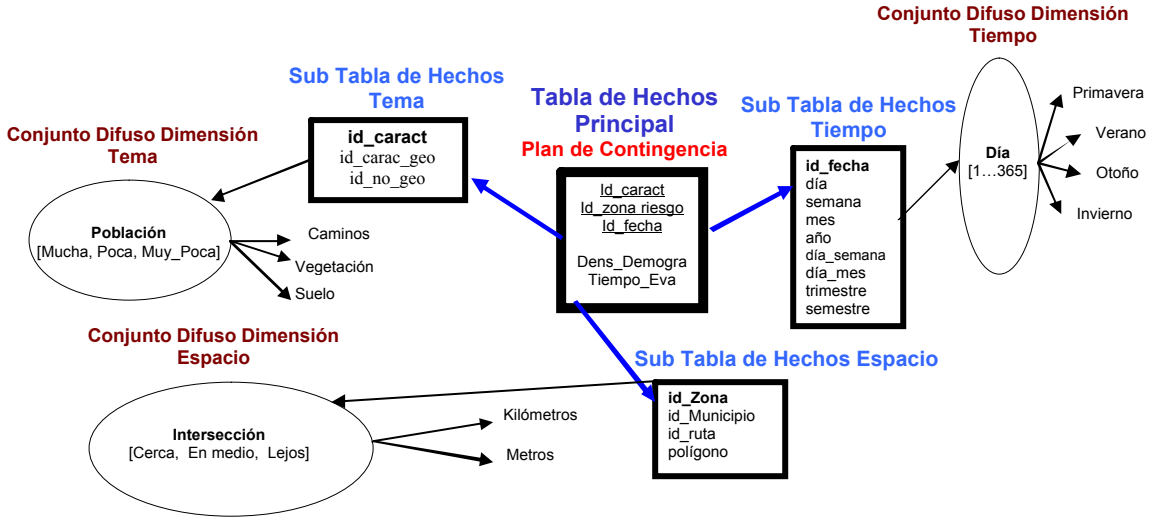


Figura 2.7 Esquema de Copo de Nieve, con los conjuntos difusos

En la figura 2.7, se observan las dimensiones consideradas hechos (principal y subtablas) con los conjuntos difusos que se obtendrán en el proceso de clasificación de cada una de las dimensiones, cabe señalar que las funciones de membresía serán definidas a partir de estos conjuntos para asignar los grados de pertenencia y plasmar estos resultados en el cubo de datos espacial difuso.

Un esquema de copo de nieve, es la transformación de un esquema de estrella basado en la tercera forma normal, las redundancias son eliminadas por la normalización de las tablas de dimensión. Para cada jerarquía de dimensión habrá una tabla por separado (tabla de dimensión). Las conexiones entre las tabla de dimensión jerárquica se realizan pasando la clave primaria a la tabla mencionada en la jerarquía. Este tipo de esquema envuelve más operaciones de reunión (join) pero hace uso de menos capacidad de almacenamiento debido a la normalización [21].

Como en nuestro caso la cantidad de información que se va a manejar exhibe alta complejidad y volumen (datos espaciales), y por lo expresado en [21] referente a la capacidad de almacenamiento que ofrece un esquema de copo de nieve, además en [22] se menciona que otra ventaja es la mejora en el desempeño de consultas debido al menor espacio usado en almacenamiento a disco, por otra parte, el almacenamiento de grandes volúmenes exige tener una buena capacidad de almacenamiento, por ello el esquema de copo de nieve es el que mejor se acopla a las tareas que se requerirán en un futuro.

2.4 Diseño Lógico Específico

En el punto anterior se propuso un esquema para la representación del Almacén de Datos Espacial Difuso, el esquema propuesto se había pensado en un inicio, considerando almacenar la información espacial en crudo, es decir tal y como los mapas la tienen almacenada en sus tablas. La desventaja de realizar esta tarea en nuestro almacén es la disminución en el desempeño del mismo así como de las herramientas OLAP que quisieran hacer uso de él, ya que la información espacial es más difícil de manipular por la naturaleza de los datos, (al tratarse de información referenciada por un sistema de coordenadas ya sea un punto, una línea o un polígono. Los almacenes de datos tienden a almacenar grandes volúmenes de información dentro de un registro);

este hecho hace que el manejo de consultas dentro del almacén de datos genere tiempos de respuesta muy largos o indefinidos. Además, los métodos de pre-agregación conocidos no pueden ser aplicados a las operaciones OLAP [6][23][26], es decir las operaciones que nos permiten generar los diferentes niveles dentro de nuestro almacén difícilmente soportarán el agrupamiento de datos espaciales por la naturaleza de los mismos.

En [6][23][24], se propone resolver este problema de los datos espaciales dentro de un almacén con tres posibilidades:

1) Colectar y almacenar los apuntadores espaciales correspondientes al objeto, pero no realizar el pre-cálculo de las medidas espaciales en el cubo de datos espacial.
2) ***Precalcular y almacenar algunas estimaciones de las medidas espaciales en el cubo de datos espacial.***
3) Selectivamente precalcular algunas medidas espaciales en un cubo de datos espacial.

Realizar el precálculo de la información que se va a guardar en el almacén de datos ayudaría a reducir el manejo de información espacial realizando consultas de acuerdo a ciertas relaciones que tengan que ver con ubicaciones espaciales que el almacén por sí solo no soporta. Una vez generada la consulta, esta regresará un conjunto de datos solución los cuales pueden ser manipulados antes de ser trasladados al almacén de datos. Este hecho está relacionado con el proceso de Extracción, Transformación y Carga de Información (ETL). Durante este proceso se pueden implementar las consultas espaciales para transformar la información referente a un sistema de coordenadas en datos relacionados a ellas, pero en términos que puedan ser agrupados o de más fácil manipulación.

El precálculo de algunas medidas espaciales en el cubo de datos espacial se puede realizar por medio de la generación de la información durante el proceso ETL [25] con herramientas que nos permitan realizar consultas espaciales y evite realizar las mismas una vez generado el Almacén.

En [25] esta tarea se propone en tres pasos:

1. Seleccionar los resultados que satisfagan mejor la consulta usando enunciados SQL proporcionados por un manejador (MySQL).
2. Cargar los resultados en el Almacén de Datos generando una nueva tabla de la dimensión espacial.
3. Realizar las operaciones OLAP (generación del cubo de datos, etc.)

El primer paso es el más importante, debido a que se pueden generar consultas espaciales en SQL complejas [25], ya que el manejo de datos espaciales requiere del conocimiento especializado del lenguaje SQL para generar agrupaciones de datos que tengan una relación de tipo espacial, (topológica, intersección, etc.). Por ello, es útil hacer uso de un Sistema de Información Geográfica que transforme estas consultas, ya que estos sistemas están especializados hacia este tipo de consultas por medio de herramientas que simplifican la generación de expresiones en lenguaje SQL. ArcGIS permite realizar algunas operaciones espaciales que pueden ser almacenadas dentro de la base de datos para realizar el paso uno propuesto en [25].

2.4.1 Transformación de la Dimensión Espacio

En la figura 2.6 se muestra el esquema de copo de nieve para todo el almacén, dentro de este almacén existe una dimensión que está clasificada por la subtabla de hechos espacial. A continuación se muestra el fragmento de la misma en la figura 2.8.

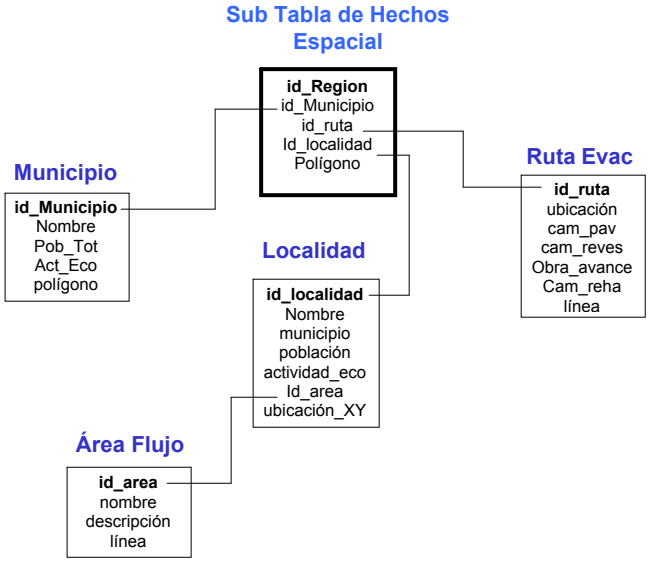


Figura 2.8 Dimensión Espacio del esquema propuesto

En la figura 2.8, *polígono, línea, ubicación_XY* se refieren a datos espaciales relacionados a un sistema de coordenadas, por lo tanto deben de ser modificados en cada una de las tablas que los contienen.

Para cada una de las tablas de la figura 2.8 se pueden generar nuevas consultas espaciales de acuerdo a la información contenida en la base de datos espacial, y así generar datos pre-agregados para ser utilizados en una consulta de tipo espacial específica a futuro [24][25]. Por ejemplo, para la tabla municipio se tiene un valor espacial (polígono), el cual tiene información de coordenadas.

Id_mun	Nombre	Pob_Tot	Act_Econo	Polígono
1	ACAJETE	351	SIEMBRA	(617350, 2096890) ,(54760, 213509), (651626, 2177004) ,(655233, 2173789), (617350,2096890) ,(54760, 213509)
2	ACATENO	423	CULTIVO	(675293, 2106590) ,(669084, 2706790), (679084, 2806293) ,(685233, 2101510), (675293, 2106590)
3	ACATLÁN	123	SIEMBRA	(545869,2096890) ,(547660, 2135079), (595239, 2273004) ,(595999, 237732), (545869,2096890)

Figura 2.9 Tabla Municipio

Como se puede observar en la figura 2.9, en el atributo Polígono, los datos son coordenadas que delimitan el polígono que corresponden al municipio que está identificado por el atributo Id_mun, pero el manejo de esta información dentro de un almacén no puede ser representada por niveles de agregación.

Por lo tanto, se realizará el precálculo de algunas medidas espaciales que estén relacionadas hacia consultas que asocien la ubicación de un elemento dentro de la base

de datos. Haciendo uso como ya se mencionó, de la herramienta ArcGis y guardando esta información dentro del almacén de datos. De esta manera se verán reflejados los hechos espaciales dentro del almacén.

Por ejemplo sea la consulta Q1: "Mostrar todos los poblados cercanos a una ruta de evacuación dada", una consulta que relaciona dos entidades espaciales, poblado (representado por un punto) y ruta de evacuación (representado por una línea), entonces, el manejador que en este caso será ArcGis, tiene que ser capaz de regresar el conjunto de datos respuesta que responda a la consulta Q1. Con ello se evita la generación de lenguaje SQL que resulte en sentencias complejas y que no se sepa si en verdad nos regresa el conjunto satisfactorio con la sentencia generada en SQL, ya que al tener la información tanto en tabla como de manera visual por medio de los mapas, los resultados de las consultas serán más fiables por simple inspección al mapa que contiene el conjunto que ArcGis ha seleccionado.

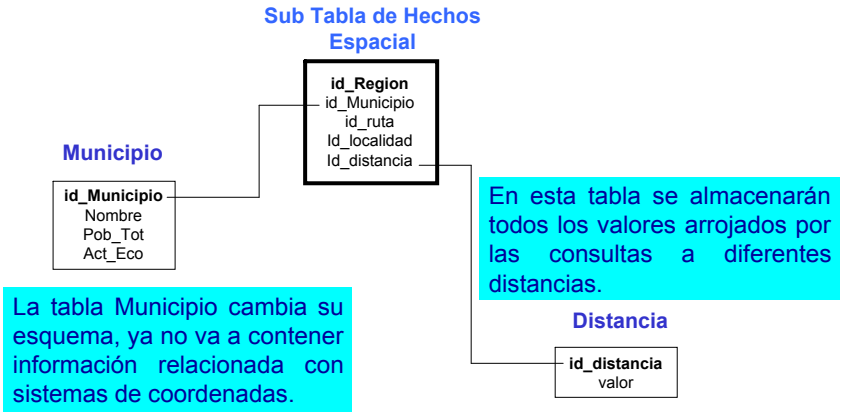


Figura 2.10 Nuevo Esquema de las tablas Municipio y Distancia

En la figura 2.10, se observan los cambios que sufre la tabla municipio con respecto a la figura 2.8 en la que se genera una nueva tabla Distancia, la cual contiene un valor de distancia que está relacionado con un identificador hacia alguna localidad y en la tabla Municipio se pierde el atributo que contenía las coordenadas del polígono que en el mapa conforman un municipio. De esta manera no se pierde la característica espacial, ya que en la distancia se refleja la relación entre dos elementos espaciales.

Por lo tanto, cuando se quiera representar información espacial dentro de un almacén, estas se pueden representar por consultas creadas en el proceso de extracción, transformación y carga del almacén por medio de consultas espaciales que relacionen dos o más entidades. Otra ventaja es que no se va a trabajar con sistemas de coordenadas dentro del almacén, sino sus representaciones por medio de las consultas creadas previamente, a estas representaciones ya se les pueden aplicar las funciones de agregación y generar el cubo de datos por ejemplo.

Más adelante, se mostrará el esquema final de copo de nieve, introduciendo los cambios necesarios para evitar el manejo de información espacial sin procesar tal y como lo muestra la figura 2.6, en la figura 2.11 se presenta el esquema propuesto, ya modificado, de la dimensión espacial, cambiando aquellos atributos de índole espacial, por su representación por medio de una consulta espacial.

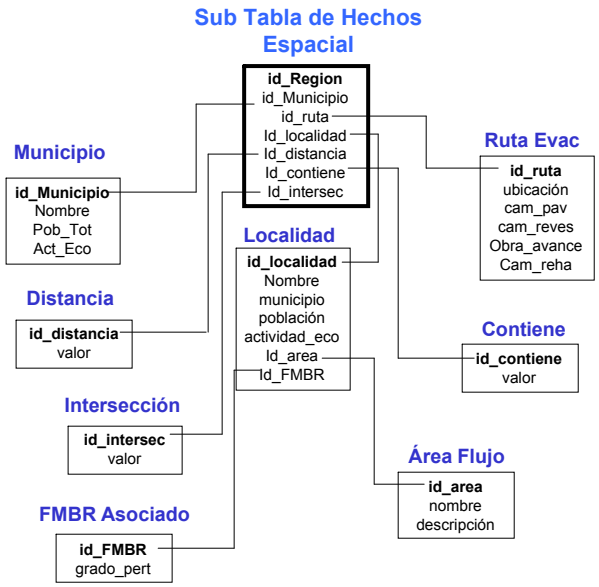


Figura 2.11 Nuevo Esquema de la dimensión Espacial

Como lo muestra la figura 2.11, Distancia, Contiene, Intersección, FMBR Asociado serán las nuevas tablas que se anexan a la dimensión Espacial, en las cuales Intersección se refiere al nivel de peligro asociado a cada población, Contiene se refiere al número de habitantes que se ubican cerca de una ruta de evacuación y FMBR Asociado se refiere a las áreas de flujo cercanas a cada poblado, de esta manera la información Espacial podrá ser representada dentro del Almacén.

2.4.2 Representación de las Funciones de Membresía en el Almacén de Datos Espacial Difuso

Una vez que se han asignado valores espaciales al almacén solo resta añadir el valor difuso a cada unos de los registros que se almacenarán en las tablas, para definir el significado de ese valor espacial dentro del almacén. Para ello, se muestra el caso de la tabla Contiene la cual, está relacionada al número de pobladores que habitan cerca de una ruta de evacuación y cuyos valores lingüísticos son Muy Pocos, Pocos y Muchos, que van en función de la cantidad de personas que viven alrededor de una ruta de evacuación tal y como lo muestra la figura 2.12:

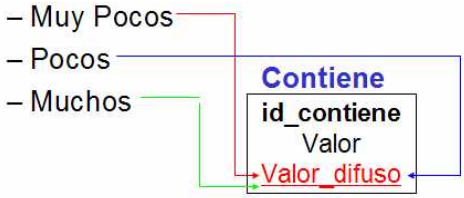
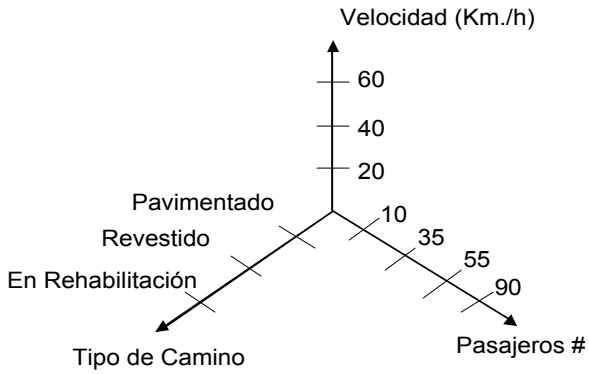


Figura 2.12 Valores lingüísticos que se le pueden asignar a la tabla Contiene

Como se puede observar en la figura 2.12, a cada valor espacial, se le asocia un valor difuso, y su valor es asignado por medio de la evaluación de una función de membresía.



	10	35	55	90
Re-60	MP	MP	P	M
Re-40	MP	P	P	M
Re-20	P	P	M	M
Pa-60	MP	MP	MP	M
Pa-40	MP	P	P	M
Pa-20	MP	P	M	M
Rh-60	P	P	M	M
Rh-40	P	M	M	M
Rh-20	M	M	M	M

MP = Muy Pocos, P = Pocos, M = Muchos

Re = Revestido, Pa = Pavimentado, Rh = Rehabilitacion

Figura 2.13 Estimación Empírica de la Variable Lingüística

La función de agregación (*f)* será definida por las combinaciones de tres variables: el tipo de camino que puede ser considerado como ruta de evacuación, la velocidad máxima permitida en los caminos y el número de pasajeros que puede llevar un transporte para realizar una evacuación. Estas variables intervienen en la evaluación del valor que le puede corresponder a una variable, los grados de membresía entre velocidad, tipo de camino y número de pasajeros, son asignados de acuerdo a los valores lingüísticos que las variables tienen.

Tabla 1 Evaluación de las variables lingüísticas

Tipo de Camino	**Pavimentado**	**Revestido**	**Rehabilitación**
μ(camino)	1	0.6	0.3
Número de Pasajeros	**De 56 a 90**	**De36 a 55**	**De 10 a 35**
μ(pasajeros)	1	0.5	0.2
Velocidad (Km./h)	**60**	**40**	**20**
μ(velocidad)	1	0.6	0.3

A partir de las estimaciones hechas para cada una de las variables (Tipo de camino, Número de pasajeros y Velocidad máxima permitida), en la tabla 1, se propone la función de agregación, como una manera de evaluar el desempeño de las tres variables de acuerdo a su integración para poder asignarle grados de pertenencia al cubo de datos por ejemplo.

$$f = \frac{ca\min o(z) * pasajeros(x)}{velocidad(y)} \tag{1}$$

Para el número de pasajeros que pueden viajar en un transporte, los valores máximos para cada una de las variables lingüísticas son "Muy Pocos" – 0.2, "Pocos" – 0.5 y "Muchos" – 1, entonces se debe de dividir el intervalo [0,1] para cada una de las variables quedando:

- *Muy Pocos* en el intervalo [0,0.2]
- *Pocos* en el intervalo [0.21,0.5]

- *Muchos* en el intervalo [0.51,1]

Con ello se puede asignar un grado de membresía a cada una de las variables lingüísticas y además permite encontrar la función de agregación para integrar a las tres variables.

Tomando {0.2, 0.5, 1} como los valores máximos de la variable pasajeros, y dividiendo el conjunto de elementos al que puede pertenecer x, encontramos para cada uno de los intervalos mediante el inverso multiplicativo [31], los coeficientes de cada subdivisión:

$$\text{Pasajeros}(x) = \begin{cases} 0.00571x & 0 \quad x \quad 35 \\ 0.009x & 35 < x \quad 55 \\ 0.011x & 55 < x \quad 90 \end{cases}$$

Se toman dichos coeficientes como en [31], para no rebasar en cada unos de los casos, los máximos valores que corresponden a los intervalos en los que se han definido ya las etiquetas lingüísticas, es decir, si x está entre 0 y 35, su valor máximo al multiplicarse con el coeficiente 0.00571, no rebasará el 0.2. El mismo caso aplica para un valor de x entre 35 y 55, el valor máximo que puede tomar es de 0.495, que está próximo a .5 que es el valor máximo para la etiqueta *Pocos*. En el último caso, para un valor de x entre 56 y 90 pasajeros, el valor es aproximado a 1 (0.99), con ello podemos evaluar la cantidad de pasajeros que puedan existir en una comunidad y asignar una etiqueta lingüística de acuerdo a los máximos definidos para dichas etiquetas.

Para el tipo de camino, no se puede obtener un valor numérico, dado que no se puede medir lo que significa un camino en Rehabilitación, Revestido o Pavimentado, esto implica que se está trabajando con una variable nominal, por ello se le asigna de manera empírica su valor, tal y como se muestra en la tabla 1.

$$\text{Camino}(z) = \begin{cases} 0.3 & z = \text{Rehabilitación} \\ 0.6 & z = \text{Revestido} \\ 1 & z = \text{Pavimentado} \end{cases}$$

Para la variable que involucra la velocidad máxima a la que se puede recorrer el camino un transporte, se procede de la misma manera que en el caso de la variable pasajeros; se toman los valores máximos de cada una de las clases que se han creado para la velocidad, y mediante el inverso multiplicativo [31] de cada uno de estos máximos, encontramos los coeficientes con los cuales nuestra variable se debe de asociar.

$$\text{Velocidad}(y) = \begin{cases} 0.01y & 0 \quad y \quad 20 \\ 0.015y & 20 < y \quad 40 \\ 0.0166y & 40 < y \quad 60 \end{cases}$$

Como en el caso de la variable *número de pasajeros*, se puede comprobar que cada uno de los coeficientes por los que se debe de multiplicar la variable ***y*** que indica la velocidad máxima a la que se puede transitar por un camino, no superan los máximos definidos en la tabla 1.

Por ejemplo, supongamos que dentro de la evaluación de estas tres variables, obtenemos el siguiente conjunto {47, Rehabilitación, 50}, procedemos a obtener cada una de las variables por separado para poder obtener $f(\mu)$:

x = 47 → 0.423
y = 50 → 0.83
Rehabilitación → 0.3

{0.423,0.3,0.83}

$$f = \frac{0.423 * 0.3}{0.83} = 0.153$$

Este valor indica la probabilidad de tener una población de 47 pasajeros en un camino en rehabilitación y que este camino se pueda recorrer a una velocidad de 50 que es cercana a la máxima permitida.

Por lo tanto, una nueva consulta necesita ser elaborada para determinar el grado de cumplimiento de la sentencia ***"E1: Recorrer un camino pavimentado a alta velocidad llevando muy pocos pasajeros".*** La tabla 1 muestra los resultados de la evaluación del enunciado, por medio de la evaluación de la función de agregación (1).

$$\min(\max[\mu(\text{camino}), \mu(\text{velocidad}), \mu(\text{pasajeros})]) = 0.3 \qquad (2)$$

Tabla 2 Evaluación de E1

	Camino	Velocidad	# pasajeros
Pa-alta-MP	1	1	0.2
Pa-media-MP	1	0.6	0.2
Pa-baja-MP	1	0.3	0.2
Pa-alta-P	1	1	0.5
Pa-media-P	1	0.6	0.5
Pa-baja-P	1	0.3	0.5
Pa-alta-M	1	1	1
Pa-media-M	1	0.6	1
Pa-baja-M	1	0.3	1

En la tabla 2 se están tomando los valores máximos como referencia en la evaluación de cada variable (a partir de los valores de la tabla 1). Tomando en cuenta las condiciones de E1, se procede a evaluar únicamente la Velocidad y el número de pasajeros, ya que E1 indica que el camino ya está pavimentado. Esto implica que se debe de omitir el tipo de camino de la evaluación de cada renglón, por que sino, el mínimo de los máximos sería siempre 1.

La tabla 1 muestra la asignación empírica de los grados de membresía a las variables lingüísticas, el grado de cumplimiento de ***E1*** fue *0.3*. Este resultado significa que es

muy poco probable que pocos pasajeros estén viajando en un camino pavimentado a la velocidad máxima permitida.

De la misma manera en que el esquema de la dimensión espacial se modificó para realizar la sustitución de datos espaciales en crudo por datos espaciales procesados por una consulta, los valores lingüísticos pueden ser asignados a cada una de las nuevas tablas, esto genera el siguiente esquema de la dimensión espacial:

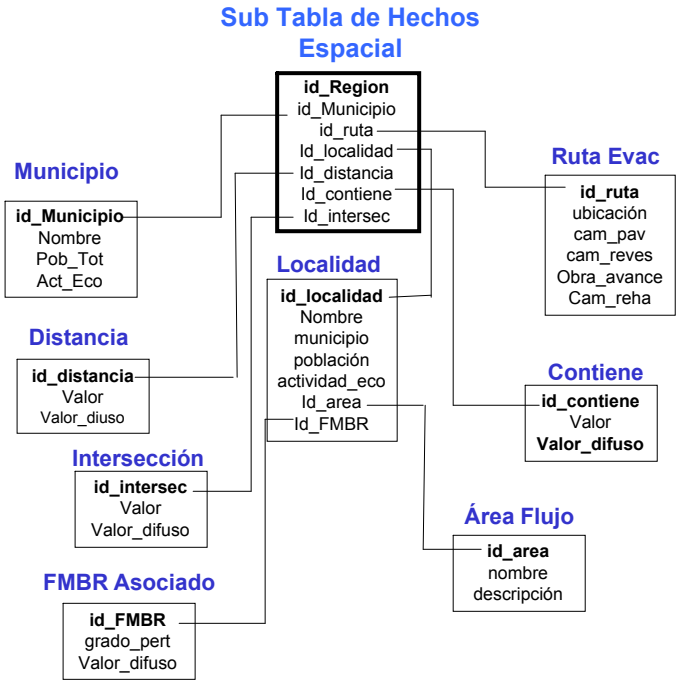


Figura 2.13 Dimensión Espacial con información espacial difusa

En la figura 2.13 se presenta el esquema final propuesto para la dimensión Espacial. Este esquema contiene la información espacial generada por medio de las consultas hacia las entidades espaciales e información acerca del significado de cada valor que contenga un atributo espacial. De esta manera los datos que sean guardados dentro del almacén de datos, reflejarán información de tipo espacial y difusa al mismo tiempo, todo esto corresponde a una sola dimensión. Se debe recordar que el almacén consta de tres dimensiones, de las cuales Tema y Espacio juegan un papel importante en el desarrollo del almacén de datos espacial difuso, todo esto se explicará más adelante cuando se presente de manera formal el caso de estudio.

En este capítulo se abordó el diseño del almacén de datos espacial difuso, tomando algunos conceptos espaciales para su representación. Se ha solucionado el problema del manejo de la información espacial, así como, de la representación de la misma en el almacén de datos, mediante el uso de herramientas en manejo de Información Geográfica como lo es ArcGIS y la selección de características espaciales que se pueden hacer con dicha herramienta.

Cabe señalar, que si se desean realizar diferentes tipos de consultas espaciales, estas se pueden realizar durante el proceso de elaboración del almacén (ETL). Se pueden generar nuevas tablas dentro del esquema propuesto y añadir estos datos al almacén, lo único que se necesita saber a priori es el tipo de datos se van a requerir en un futuro para

realizar las consultas pertinentes y armar un almacén que esté enfocado a representar toda esta información para cualquier tipo de consulta o uso que se le pueda hacer.

CAPÍTULO 3

IMPLEMENTACIÓN DEL CUBO DE DATOS

3.1 Herramientas de Desarrollo ó Diseño Físico

3.1.1 Mondrian OLAP

La herramienta OLAP Mondrian, será la proveedora de nuestro esquema ROLAP, la cual está implementada en el lenguaje Java y es la encargada de realizar el proceso de extracción, transformación y carga de los datos, ya que cuenta con varias capas que soportan estas tareas, y que a continuación se explican más a detalle.

Un Sistema Mondrian OLAP consiste de cuatro capas; trabajando desde la presentación del usuario final a las entrañas de los datos, estas son: la capa de presentación, la capa dimensional, la capa de estrella y la capa de almacenamiento (ver la figura 3.1). La capa de presentación determina lo que el usuario final ve en su monitor, y la forma de interactuar para realizar nuevas consultas. Hay muchas formas para presentar los conjuntos de datos multidimensionales: tablas de pivote (una versión interactiva), rebanada, líneas y barras. Estas formas de presentación tienen en común la gramática multidimensional de las dimensiones, medidas y celdas en las cuales la capa de presentación realiza la consulta, y el servidor OLAP regresa la respuesta [27].

La segunda capa es la capa dimensional, la cual hace un análisis sintáctico, válida y ejecuta las consultas MDX. Una consulta es evaluada en múltiples fases. Los ejes son calculados primero, después los valores de las celdas dentro de los ejes. La capa dimensional envía una petición de celda a la capa de agregación por lotes. Un transformador de la consulta permite a la aplicación manipular las consultas existentes, en vez de construir una declaración MDX de la nada para cada petición. Los metadatos describen el modelo dimensional, y cómo se mapean dentro del modelo relacional [27].

La tercera capa es la capa de estrella, y es responsable de mantener un cache agregado. Una agregación es un conjunto de valores de medidas (celdas) en memoria, calificados por un conjunto de valores de dimensión columna. La capa dimensional envía peticiones para un conjunto de celdas. Si las celdas requeridas no están en el cache, el administrador de agregación envía una petición a la capa de almacenamiento [27].

La capa de almacenamiento es un Sistema Administrador de Bases de Datos Relacionales. Es responsable de proporcionar los datos de celda agregados y los miembros de las tablas dimensión. Todos estos componentes pueden existir en una sola máquina, o pueden estar distribuidos entre varias máquinas. Las capas 2 y 3, las cuales comprenden el servidor Mondrian, deben estar en la misma máquina. La capa de almacenado puede estar en otra máquina, accesada por medio de una conexión remota JDBC [27].

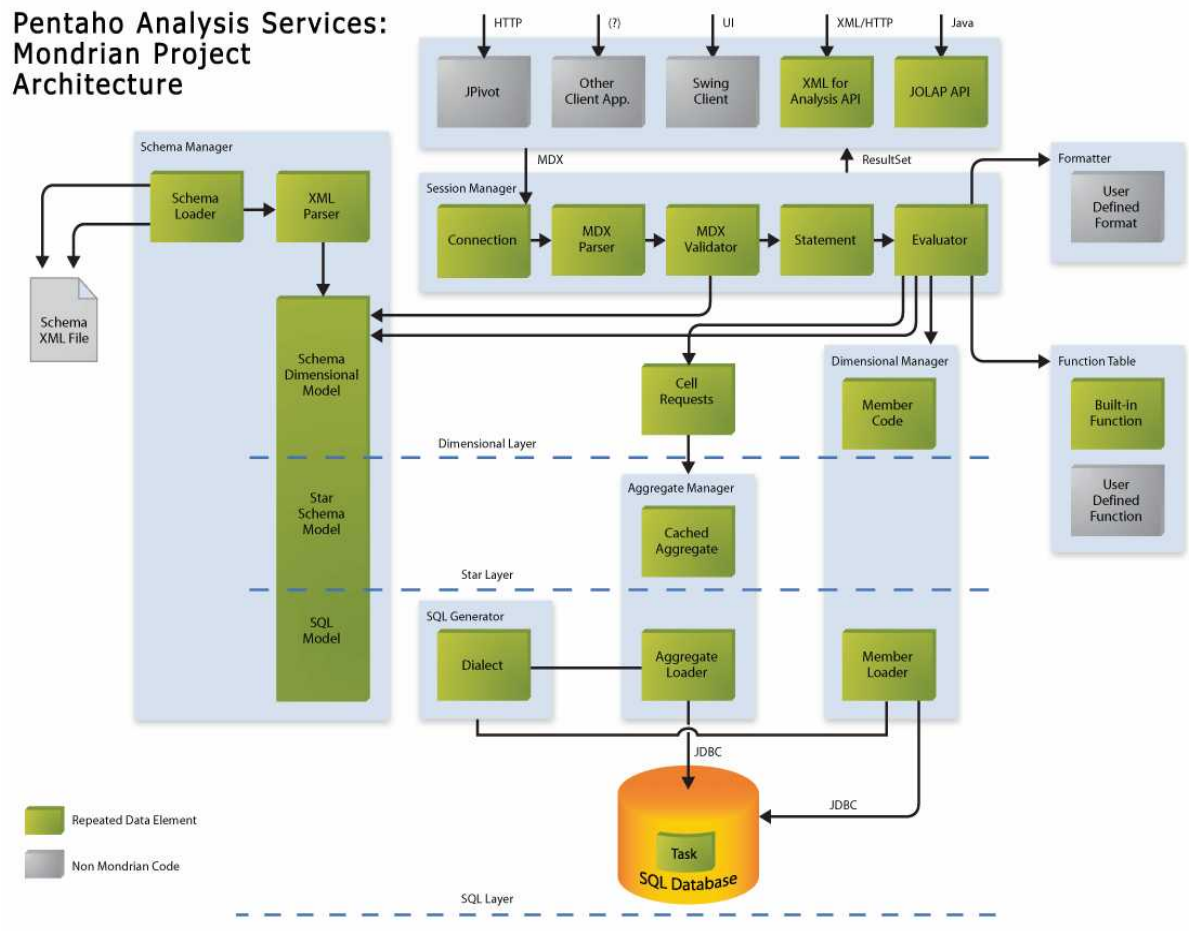


Figura 3.1 Capas de la Arquitectura de Mondrian. [27]

Los Servidores OLAP son generalmente categorizados de acuerdo a como almacenan sus datos:

Un servidor MOLAP, almacena todos sus datos en disco en estructuras optimizadas para acceso multidimensional. Típicamente, los datos son almacenados en grandes arreglos, requiriendo solo 4 o 8 bytes por un valor en la celda.

Un servidor ROLAP, almacena sus datos en una base de datos relacional. Cada fila en una tabla de hechos tiene una columna para cada dimensión y medida [27].

Tres tipos de datos necesitan ser almacenados: los datos de la tabla de hechos (los registros transaccionales), agregaciones y dimensiones.

Las bases de datos MOLAP almacenan los datos que son hechos en un formato multidimensional, pero si hay más que unas cuantas dimensiones, estos datos estarán dispersos, y el formato multidimensional no tendrá un buen desempeño. Un sistema HOLAP (OLAP híbrido) solucionaría este problema dejando los datos mas granulados en la base de datos relacional, pero almacenaría las agregaciones en un formato multidimensional [27].

Las agregaciones precalculadas son necesarias para grandes conjuntos de datos, de otra forma ciertas consultas no podrían ser respondidas sin hacer la lectura de los contenidos enteros de la tabla de hechos. Las agregaciones MOLAP son siempre una imagen de la

estructura de los datos en memoria. En algunos sistemas ROLAP son explícitamente administrados por el servidor OLAP; en otros sistemas, las tablas son declaradas como vistas materializadas, y son implícitamente usadas cuando el servidor OLAP genera una consulta con la combinación correcta de las columnas en la cláusula group by [27].

El componente final de la estrategia de agregación es el caché. El caché mantiene agregaciones precalculadas en memoria de forma que consultas subsecuentes puedan acceder a los valores de las celdas sin ir al disco. Si el caché mantiene el conjunto de datos requeridos en un nivel muy bajo de agregación, entonces puede calcular el conjunto de datos necesarios para la operación de Roll up [27].

El caché es posiblemente la parte más importante de la estrategia de agregación ya que es adaptativa. Es difícil elegir un conjunto de agregaciones para precalcular con gran velocidad del sistema sin utilizar una gran cantidad de disco, particularmente aquellas con una gran dimensionalidad o si los usuarios están ingresando consultas impredecibles. En un sistema donde un dato está cambiando en tiempo real, es impráctico mantener agregaciones precalculadas. Un tamaño razonable de caché puede permitir a un sistema desempeñarse adecuadamente al enfrentar consultas impredecibles, con pocas agregaciones precalculadas o sin ellas [27].

Las estrategias de agregación de Mondrian se presentan en la figura 3.2:

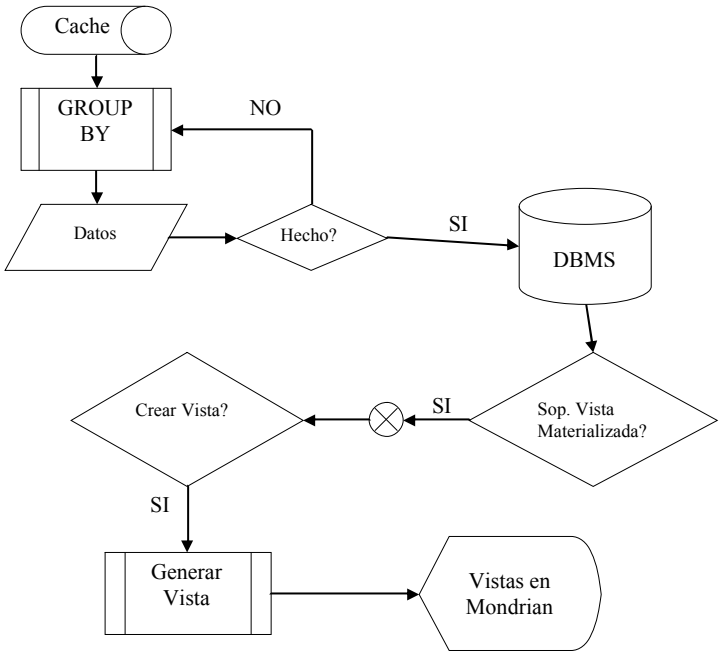


Figura 3.2 Diagrama de Secuencia de una Agregación en Mondrian

- Un dato Hecho es almacenado en el Sistema Manejador de Bases de Datos.
- Lee los datos agregados dentro del cache ingresando consultas group by.
- Si el Sistema Manejador soporta las vistas materializadas, y el administrador de la base de datos elige crear vistas materializadas para una agregación en particular, entonces Mondrian los usará de manera implícita. Idealmente, el

manejador de agregaciones Mondrian debería estar pendiente de que estas vistas materializadas existan y que las agregaciones particulares son fáciles de calcular.

La idea general es delegar a la base de datos el conjunto de tareas que ésta sabe desempeñar. Esto coloca carga adicional a la base de datos, pero una vez que estas características son añadidas a la base de datos, todos los clientes de la base de datos se beneficiarán de ellas. El almacenamiento multidimensional reduciría entradas, salidas y resultados en operaciones más rápidas en algunas circunstancias [27].

Un efecto secundario maravilloso es que por el hecho de que Modrian no requiere de almacenamiento para él, este puede ser instalado añadiéndole un archivo JAR al classpath, ser levantado y poder ejecutarlo de inmediato. Ya que no hay conjuntos de datos redundantes para manejar, el proceso de carga de datos es más fácil, y Mondrian es muy apropiado para realizar el OLAP sobre los conjuntos de datos que cambian en tiempo real [27].

Interfaz de Programación de Aplicaciones (API- Application Programming Interface).
Mondrian proporciona un API para aplicaciones de cliente que ejecuten consultas. Cualquiera que haya utilizado el JDBC encontrará familiar esta API, la principal diferencia es el lenguaje de consulta, ya que Mondrian hace uso de un lenguaje llamado MDX (Multi-Dimentional eXpressions) para especificar las consultas, JDBC usaría SQL. MDX será descrito más adelante en nuestro caso de estudio.

El siguiente fragmento de código Java de la figura 3.2 conecta a Mondrian con el DBMS, ejecuta una consulta e imprime el resultado:

```
import mondrian.olap.*;
import java.io.PrintWriter;

Connection connection = DriverManager.getConnection(
"Provider=mondrian;" +
"Jdbc=jdbc:odbc:MondrianFoodMart;" +
"Catalog=/WEB-INF/FoodMart.xml;",
null,
false);
Query query = connection.parseQuery(
"SELECT {[Measures].[Unit Sales], [Measures].[Store Sales]} on
columns," +
" {[Product].children} on rows " +
"FROM [Sales] " +
"WHERE ([Time].[1997].[Q1], [Store].[CA].[San Francisco])");
Result result = connection.execute(query);
result.print(new PrintWriter(System.out));
```

Figura 3.3 Código Java para conectar con Modrian y realizar una consulta.

Una conexión es creada por medio del DriverManager, en una forma similar a JDBC. Una consulta es análoga a un enunciado JDBC, y es creada analizando una cadena MDX. Un resultado es análogo a un ResultSet JDBC, ya que se está tratando con datos multidimensionales, que consiste de ejes y celdas más que de renglones y columnas. Además, ya que OLAP está enfocado para la exploración de los datos, se puede modificar el árbol de análisis contenido en una consulta para operaciones como drillDown, ordenamiento, etc., y ejecutar la consulta. La API también presenta el

esquema de la base de datos como un conjunto de objetos: Esquema, Cubo, Dimensión, Jerarquía, Nivel, Miembros [27].

Para obedecer con los estándares emergentes, se ha agregado a Mondrian dos APIs :

JOLAP.- Es un estándar emergente de los procesos JSR, y que llegó a ser parte de J2EE algún tiempo en 2003 [27].
XML para análisis.- Es un estándar para acceder al servidor OLAP por medio del SOAP (Protocolo de Acceso a Objetos Simple). Esto le permitirá a un componente no-Java como Microsoft Excel ejecutar las consultas en lugar de Mondrian [27].

3.1.2 Pentaho Cube Designer

Esta Herramienta, facilita la creación de la estructura de nuestro almacén así como del cubo de datos que se quiere representar mediante las consultas en el lenguaje MDX, haciéndolo de manera visual y más amigable al usuario común que esta interesado en realizar tareas que involucran la creación de un almacén y un cubo de datos para el análisis.

La ventaja que tiene esta herramienta además de las antes mencionadas es que se presenta como una aplicación independiente [28], es decir que no necesita de Mondrian, por ejemplo, para trabajar con un almacén y representarlo, algunos de los inconvenientes es que se deben de tener instalados en la máquina donde se ejecuta esta herramienta, el manejador JDBC en cualquiera de sus versiones y la máquina virtual de java, esta última puede generar demasiados dolores de cabeza por la gama de programas que hacen uso de diferentes máquinas virtuales.

Cube Designer soporta los siguientes tipos de esquemas para la representación del cubo:

- Esquema de una sola tabla (Las medidas y las dimensiones son extraídas de una sola tabla).
- Esquema de Estrella (Se tienen una tabla de hecho y múltiples tablas de dimensiones).
- Esquema de Copo de nieve (Una Tabla de Hechos y varias tablas de dimensión organizadas de manera jerárquica).

Por lo tanto la herramienta soporta el esquema propuesto en el capítulo anterior. Cabe mencionar que para la creación de relaciones entre las tablas, se deben generar Joins de manera manual [28] en el panel de diseño de consultas siguiendo las siguientes reglas [28]:

- La tabla de hechos siempre debe de estar del lado izquierdo de la pantalla
- Las tablas de dimensión siempre deben de estar del lado derecho de la pantalla
- Se debe de comenzar por nivel más alto de las tablas de dimensión y después por los niveles más bajos del lado derecho (Ejemplo: Hecho → Año → Trimestre → Mes → Día).

Pantalla de Inicio

Cuando se carga Cube Designer, se despliega la pantalla inicial como se muestra en la figura 3.4. Se debe cargar un nuevo esquema de Cubo dentro del menú Archivo en la opción "New Cube Schema".

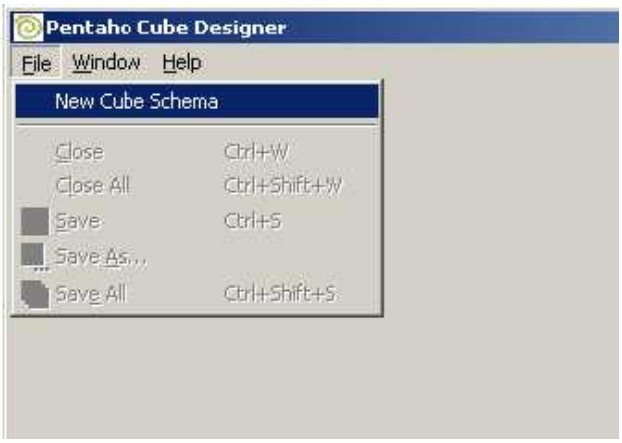


Figura 3.4 Pantalla de inicio de Cube Designer

Especificación del nombre del Cubo

La creación del esquema para Mondrian y del cubo consta de 6 pasos, el primero es especificar el nombre del cubo para que este sea reconocido por el esquema de Mondrian y el cubo de datos, cabe aclarar que el esquema de Mondrian no es más que un archivo con extensión XML, en donde se organizan y se definen todas las jerarquías que el almacén y el cubo van a tener. La figura 3.5 muestra la primera de las 6 pantallas en la creación de un cubo.

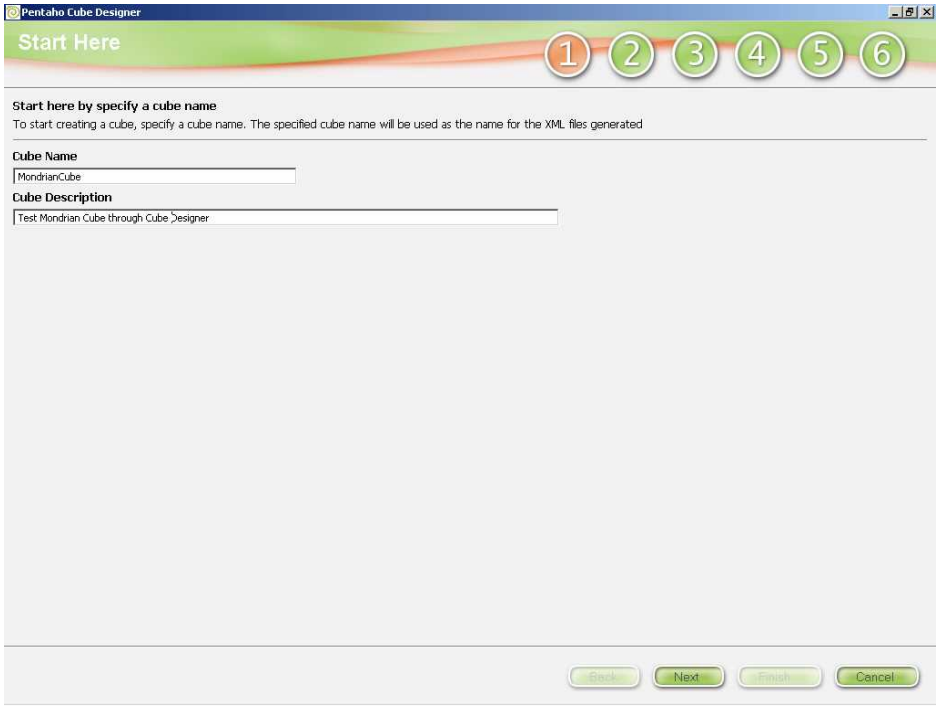


Figura 3.5 Primer Paso en la construcción de un cubo

En la figura 3.5 se presenta el inicio de la creación de un cubo mediante Cube Designer, el proceso consta de seis pasos, en el primero se asigna el nombre del cubo y una pequeña descripción del mismo, lo cual genera un archivo txt dentro del directorio de creación del cubo, para tener un conocimiento global de lo que consta ese cubo creado.

Generar un origen de datos o seleccionar uno existente

En el paso 2, se genera la cadena de conexión de la que el almacén y el cubo harán uso en un futuro para generar las consultas hacia el almacén y que estas sean representadas en el cubo. En este paso debemos dar un nombre de origen de datos, el manejador que se va a utilizar, la cadena de conexión a usar junto con el nombre de la base de datos alojada en ese manejador, el nombre del usuario que creó o generó esa base de datos o que tiene los permisos para lectura y escritura de la base, así como el password de ese usuario, esto se ejemplifica mejor en la figura 3.6 donde se genera el origen de datos.

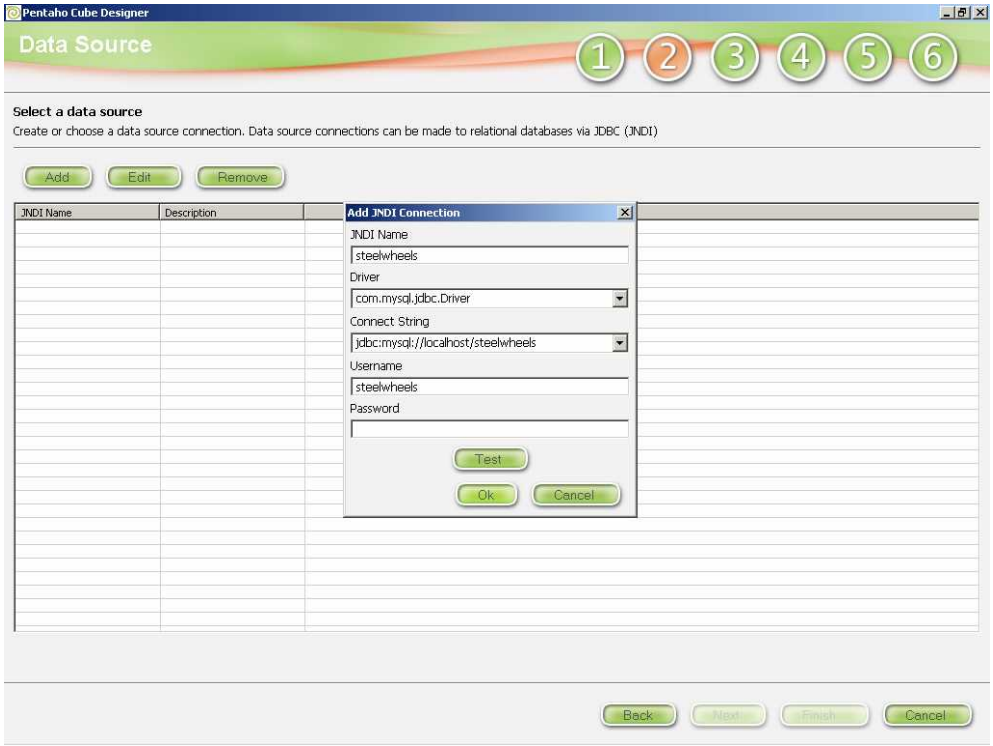


Figura 3.6 Paso 2, creación de la Fuente de Datos

Seleccionar el Origen de Datos

El paso 3 va de la mano con el paso dos, ya que dependiendo del origen de datos que se desee ocupar y que ya se tenga creado, será como elementos en la lista se desplieguen, en este paso lo único que se debe de hacer es seleccionar de la lista origen de datos, cual se quiere utilizar, como en la figura 3.6 se creó un origen de datos llamado steelwheels. En la figura 3.7 se muestra este nombre, una vez elegido el origen se da botón Siguiente.

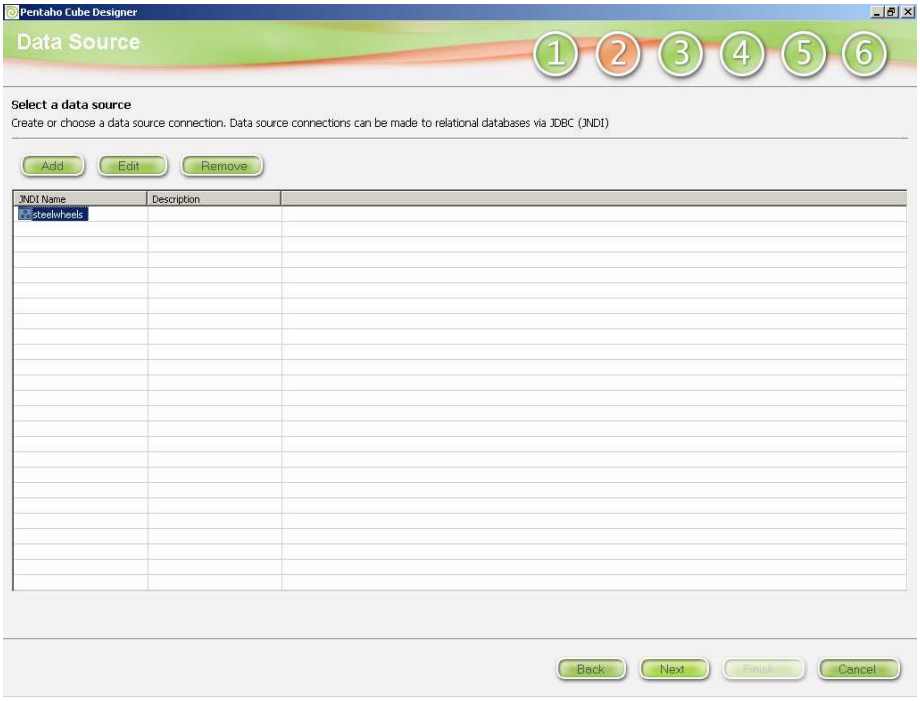


Figura 3.7 Paso 3 elegir el origen de datos a usar

Mapeo de Tablas

En este paso se construye de acuerdo al esquema que se haya propuesto y haciendo uso del panel de diseño, la representación lógica del almacén, como ya se mencionó anteriormente, este diseñador de cubos soporta el esquema de estrella y el de copo de nieve. De manera similar a como en otros manejadores de interfaz visual (Access, SQL Server, etc) se pueden generar las relaciones entre las diferentes claves definidas (ver figura 3.8) en las tablas, Cube Designer permite el mismo manejo para el diseño del esquema.

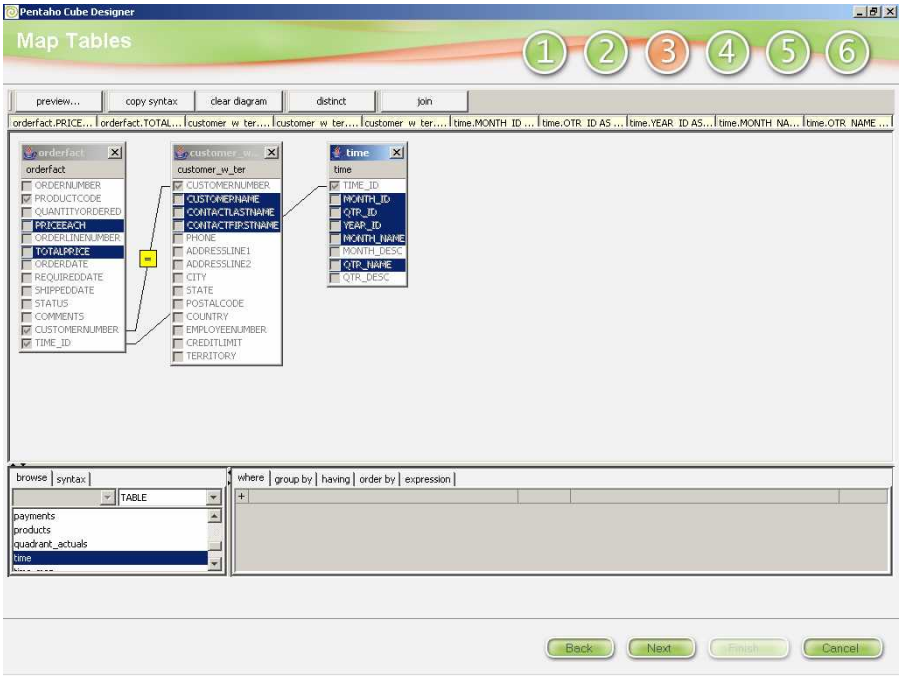


Figura 3.8 Mapeo de las Tablas en la elaboración del esquema

Crear las medidas y el conjunto de atributos

En el paso 4 se crean las medidas que se van a utilizar en la tabla de hechos que servirán para realizar el conocimiento que se quiere obtener de la información guardada en el almacén de datos. Estas medidas dependerán de las tablas que se hayan obtenido de la base en el paso 3, también en se les aplicará a las medidas una función de agregación, (suma, conteo, promedio, etc.), así como un formato y la etiqueta a desplegar dentro de la representación del Cubo. Un ejemplo del paso 4 lo muestra la figura 3.9.

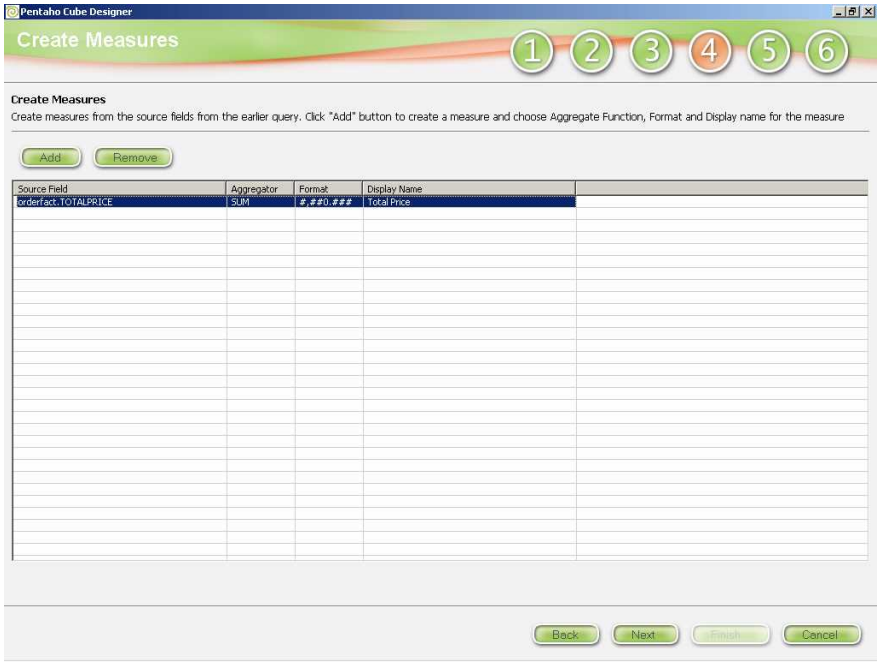


Figura 3.9 Paso 4, creación de las medidas

Creación de las Dimensiones

Las dimensiones que se crean son desplegadas como un árbol, aquí el elemento que está en la raíz del árbol es la Dimensión, el segundo elemento es la jerarquía y los subsecuentes elementos son los niveles de la dimensión como lo muestra la figura 3.10.

Dimensión
Jerarquía
Nivel 1
Nivel 2
. . .
Nivel n

Figura 3.10 Representación de una dimensión como un árbol

El panel que se encuentra en la figura 3.11, muestra las propiedades de la Dimensión, Jerarquía o Niveles seleccionado(a). Todos estos elementos consisten de atributos mezclados que son representados en las propiedades del panel situado en la parte inferior de la figura 3.11.

Un nivel puede tener propiedades que se ajusten a un requerimiento. Esto se logra usando el botón "Add Property". Una vez que se haya elegido la propiedad, se selecciona el campo de origen deseado y se especifica un nombre para la propiedad (El valor de la columna).

Para crear una dimensión, se debe seleccionar un campo origen en el panel del lado izquierdo (figura 3.11) y presionar el botón "Agregar Nueva Dimensión". Para agregar niveles a la dimensión, se debe seleccionar un campo del lado izquierdo de la figura 3.11, seleccionar el nivel superior del lado derecho y elegir el botón "→".

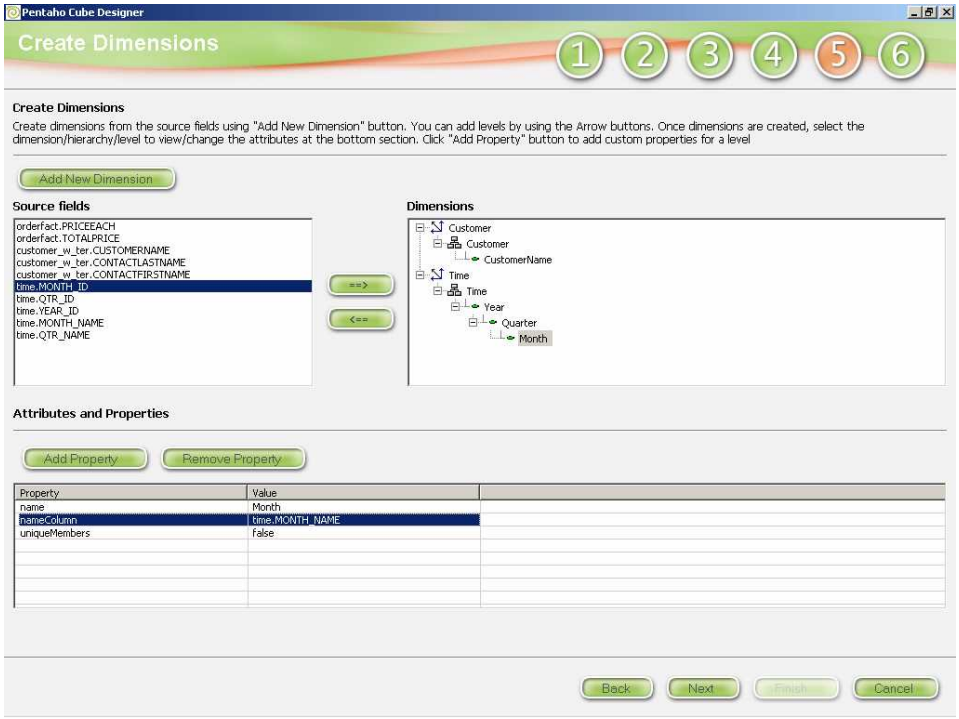


Figura 3.11 Creación de las dimensiones y niveles

Visualización de resultados

Una vez que se hayan creado todas las dimensiones deseadas, el último paso consta de una serie de pasos para poder visualizar la información que se ha creado en los 5 pasos anteriores. De estos pasos el más importante es la visualización del esquema mediante el archivo XML (que se explicará a mejor en la implementación del Almacén), ya que este archivo es el que tiene todas las definiciones acerca de la representación del Cubo en Mondrian. Lo importante de Cube Designer es precisamente la generación de esta definición, ya que sin esta herramienta se tendría que diseñar a mano, sin una visualización (casi a ciegas). El esquema del que estaría compuesto un cubo Mondrian y en este desarrollo se podrían generar errores que no permitirían elaborar un cubo de datos de manera correcta, es decir con errores que no dejarían representar las dimensiones, jerarquías o niveles de acuerdo a lo planeado en el diseño del cubo.

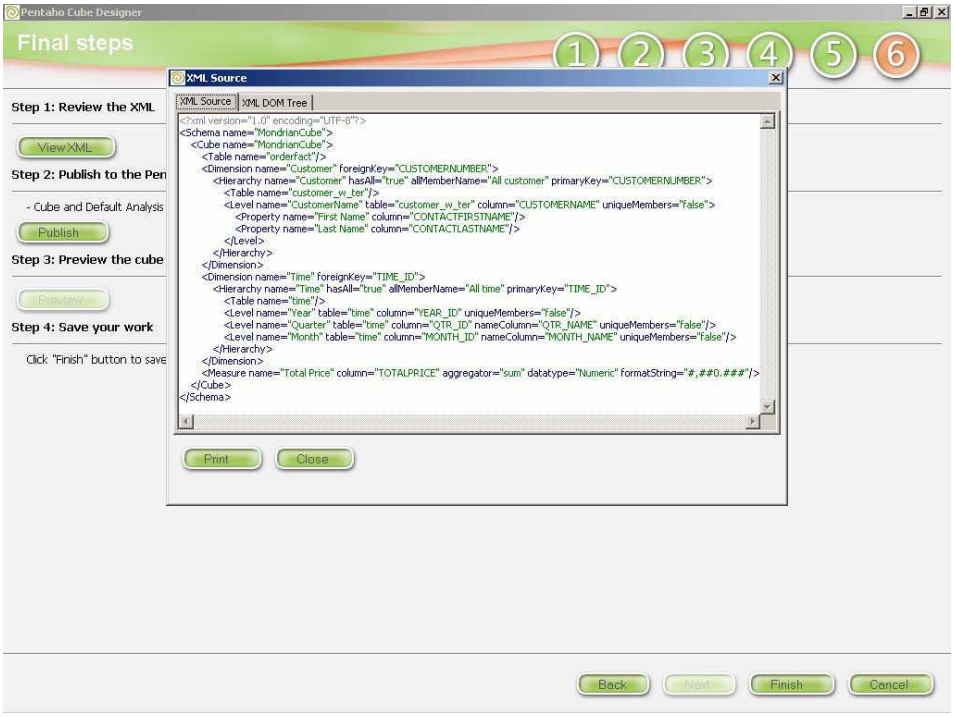


Figura 3.12 Paso 6, representación del esquema mediante XML

En la figura 3.12 se observa el archivo XML que forma el diseño del esquema, necesario para que Mondrian represente un cubo de datos, sin las definiciones que el archivo XML proporciona, no sería posible representar algún cubo mediante Mondrian, ya que este forma la estructura lógica del Almacén y hace referencia hacia las tablas de la base de datos que se requieren para formar el almacén, así como la definición de la cadena de conexión a usar con su manejador.

Para que el almacén pueda ser representado por el cubo de datos, se debe de publicar esta información dentro del motor de Pentaho, al cual se le debe de especificar la ubicación de publicación, el URL para su publicación, la contraseña de publicación, el nombre del servidor y su contraseña, tal y como se muestra en la figura 3.13.

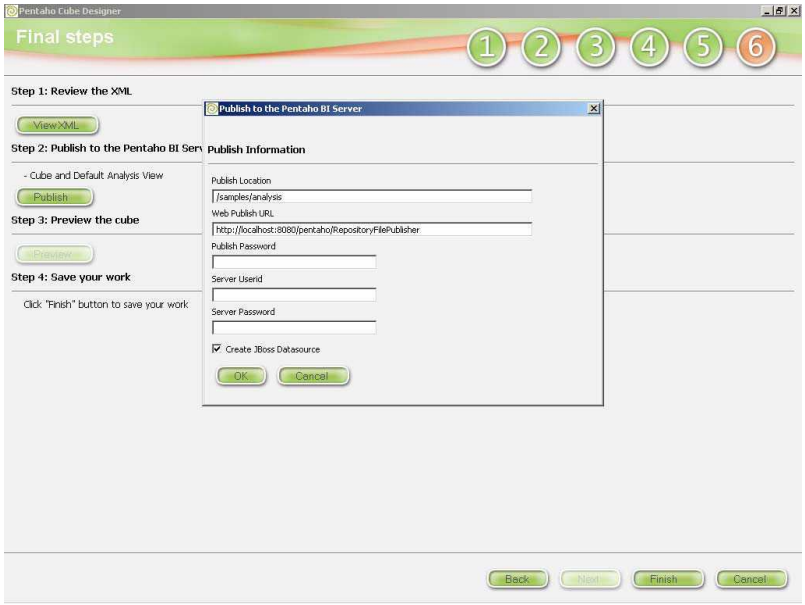


Figura 3.13 Publicación del Almacén creado

Una vez que se ha publicado el almacén se puede previsualizar dentro de este último paso, o simplemente se puede acceder a la dirección URL, con la que fue publicado, de cualquiera de las dos formas es posible visualizar el almacén representado por el cubo. Como lo muestra la figura 3.14, la previsualización se realiza oprimiendo el botón "Preview".

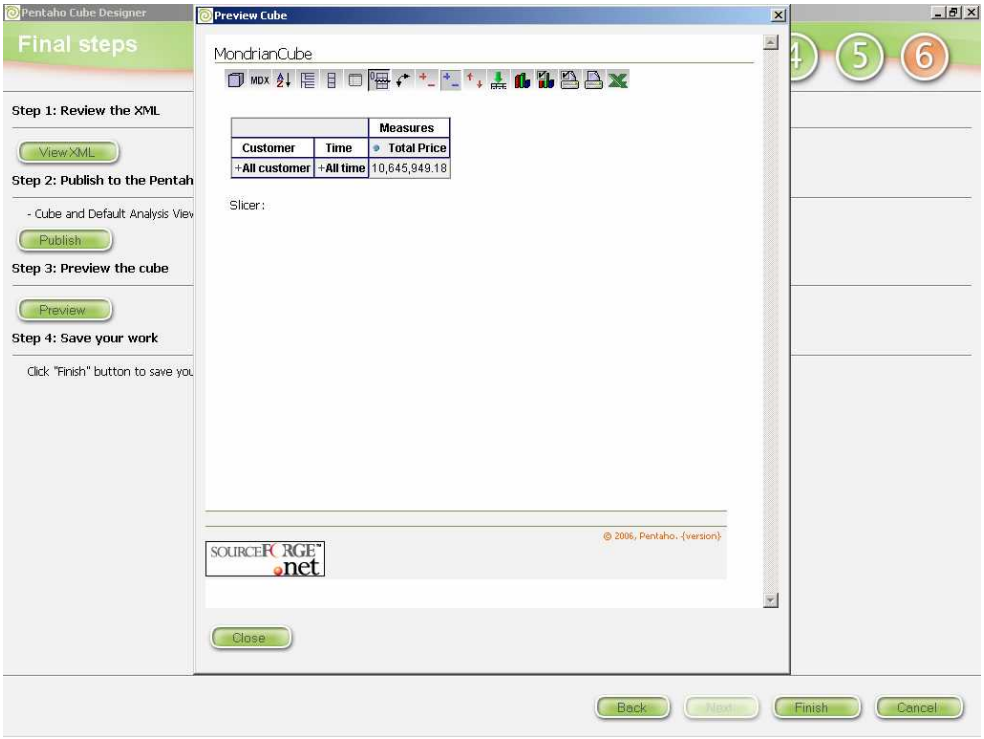


Figura 3.14 Previsualización del Cubo de datos

Finalmente, se tiene que salvar en algún directorio para usos posteriores, con ello se evita estar generando el mismo almacén varias veces dentro del diseñador de cubos. En sí lo que se guarda es el archivo XML, un archivo que contiene las especificaciones de la generación del cubo y el archivo que contiene los comentarios o propiedades del almacén creados en el paso uno.

3.1.3 Apache Tomcat

Apache Tomcat es un contenedor de servidores que es utilizado en la implementación de referencia para las tecnologías Java Servlet y Java Server Pages. Las especificaciones para estas tecnologías son desarrolladas por Sun bajo el Proceso Java Community [29].

Esta herramienta es la encargada de proporcionar comunicación entre las peticiones de usuario y las respuestas del servidor, para la interacción entre las diferentas páginas jsp y la visualización de los resultados, todo esto bajo la licencia de Apache Software [29].

Por lo tanto, si se quiere visualizar un cubo de datos creado ya sea usando el diseñador de cubo (Cube Designer) o desarrollando de manera independiente el esquema en un archivo XML y su representación mediante consultas en lenguaje MDX, se debe primero inicializar el servidor Apache para levantar el servicio de comunicación cliente-servidor y todas las aplicaciones que requieren este tipo de servicio. Otra ventaja que

tiene este tipo de servidor es que al ser un programa de distribución libre, (Open Source) puede ser usado en plataformas como Windows y Linux.

3.2 Implementación del cubo usando las herramientas

Una vez, que se ha explicado a grandes rasgos, las diversas herramientas que contribuyen a la generación del almacén para su posterior representación mediante un cubo, el siguiente y último paso en nuestro diseño de Almacén de Datos, es presentar la implementación para nuestro caso de estudio que es la finalidad de la tesis propuesta. Para ello se define lo que en [27] se denomina esquema Mondrian, en este esquema se definen los conceptos de medidas, dimensiones, jerarquías y niveles.

Un esquema se define a través de un archivo XML, el cual va definir las estructuras que van a intervenir en el almacén de datos, así como, en la generación de los cubos de datos. Una medida es el componente que uno está interesado en obtener o calcular, las dimensiones son atributos en los que se pueden dividir las medidas, las jerarquías denota la forma en que está organizada una dimensión y los niveles reflejan las jerarquías que se definen en una dimensión [27].

Para el caso de estudio que en este trabajo se propone, una dimensión a usar se presenta a continuación en la figura 3.15, con la cual se explicarán los elementos dimensión, jerarquía, nivel y propiedad.

```
<Dimension name="Espacio">
<Hierarchy hasAll="true" primaryKey="ID_LOC"
allMemberName="Todos" defaultMember="Todos">
  <Table name="Puebla"/>
  <Level name="Estado" column="NOMENT"
uniqueMembers="false"/>
  <Level name="Municipio" column="NOMMUN"
uniqueMembers="false"/>
  <Level name="Localidad" column="NOMLOC"
uniqueMembers="true"/>
  <Property name="Latitud" column="LAT"/>
  <Property name="Longitud" column="LONG"/>
  <Property name="Coordenada X" column="X"/>
  <Property name="Coordenada Y" column="Y"/>
</Hierarchy>
</Dimension>
```

Figura 2.10 Una Dimensión del Almacén de Datos propuesto

En la figura 2.10 se observa código XML en el que se define parte de un esquema Mondrian, con el cual podemos representar un almacén de datos y diversos cubos para realizar análisis de los datos. En este código se define la dimensión “Espacio” mediante las etiquetas “<Dimension name>” y “</Dimension>” (recordar que en XML, cada etiqueta que sea de apertura debe de corresponderle una de cierre, por sintaxis del lenguaje XML).

También se define el elemento jerarquía con la etiqueta “<Hierarchy>” y “</Hierarchy>” en la cual se definen cuatro cosas, la primera “ hasAll=”True” ”, es si el nivel superior de la jerarquía será omitido o no, si ponemos esta condición en falso, el nivel superior será omitido en caso contrario se presentará éste, la segunda “ primaryKey="ID_LOC" “, hace referencia a la clave primaria en la tabla de donde se

extraerán los datos que serán separados por niveles, la tercera y cuarta“ allMemberName="Todos" defaultMember="Todos” ” son formas de nombrar la representación gráfica de la jerarquía, es decir, la leyenda que aparecerá en la interfaz gráfica.

El elemento “ <Table name=”Puebla”/> “ indica la tabla de la base de datos a utilizar para nuestra dimensión. Los niveles en los que se organiza la jerarquía definida en la dimensión “Espacio”, son generados mediante la etiqueta “<Level/>”, en esta sección se debe de definir el nombre del nivel “ name="Estado" “, el atributo de la base de datos que será recuperado y clasificado de acuerdo al nivel en el que esté “ column="NOMENT" “ y si existen o no elementos repetidos dentro del nivel “ uniqueMembers="false" “.

Los niveles de una jerarquía pueden tener propiedades que también pueden ser utilizadas en las consultas para obtener información clasificada. Las propiedades son definidas mediante la etiqueta “<Property/>“, los atributos de esta etiqueta serán, el nombre de la propiedad a mostrar “ name="Latitud" “ y la columna en la tabla de la base de datos que contiene dicha información “ column="LAT" “.

Otro ejemplo en cube designer y aprovechando las características de la interfaz que nos presenta se muestra en la figura 3.16, en donde por un lado se muestra la forma en la que se organiza de manera jerárquica la dimensión a mostrar en el cubo y por otro la estructura que guarda cada uno de los elementos

```
<?xml version="1.0" encoding="UTF-8"?>
<Schema name="Prueba">
  <Cube name="Piroclasticos">
    <Table name="loc_piroclas_alto"/>
    <Dimension name="loc_piroclas_alto.CLAVE">
      <Hierarchy name="loc_piroclas_alto.CLAVE" hasAll="true"  allMemberName="All
loc_piroclas_alto.clave">
        <Table name="loc_piroclas_alto"/>
            <Level name="loc_piroclas_alto.CLAVE" table="loc_piroclas_alto
column="CLAVE" uniqueMembers="false"/>
        <Level name="loc_piroclas_alto.NOMLOC" table="loc_piroclas_alto
column="NOMLOC" uniqueMembers="false"/>
        <Level name="loc_piroclas_alto.POBLACION" table="loc_piroclas_alto
column="POBLACION" uniqueMembers="false"/>
          <Level name="loc_piroclas_alto.PIROCLASTICO" table="loc_piroclas_alto
column="PIROCLASTICO" uniqueMembers="false"/>
      </Hierarchy>
    </Dimension>
      <Measure name="SUM of loc_piroclas_alto.CLAVE" column="POBLACION
aggregator="sum" datatype="String" formatString="#,##0"/>
    <Measure name="SUM of loc_piroclas_alto.CLAVE" column="PIROCLASTICO
aggregator="count" datatype="String" />
  </Cube>
</Schema>
```

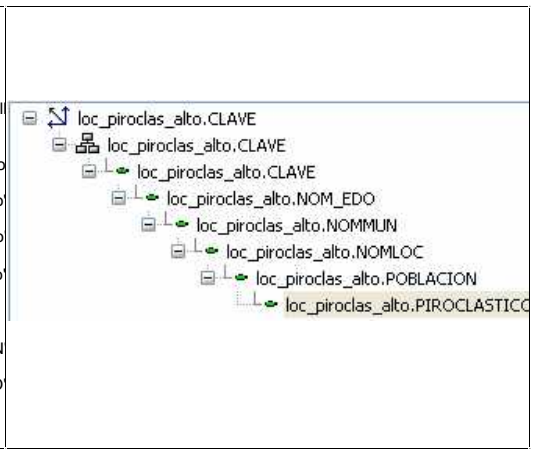


Figura 3.16 Esquema y Forma de Jerarquías en la Dimensión Piroclástico

En la figura 3.16 se observa el esquema en XML a la izquierda y del lado derecho la manera en que se organiza la jerarquía que será representada en el cubo, la definición de este esquema es necesaria para que sea visualizado el cubo tanto en Mondrian como en Cube Designer, a continuación se presenta una parte del cubo en la que se muestra el tipo de peligro asociado a una cantidad de pobladores de una localidad (Tlepatlacita) la cual se encuentra en la zona que corresponde a un peligro de tipo “Alto”.

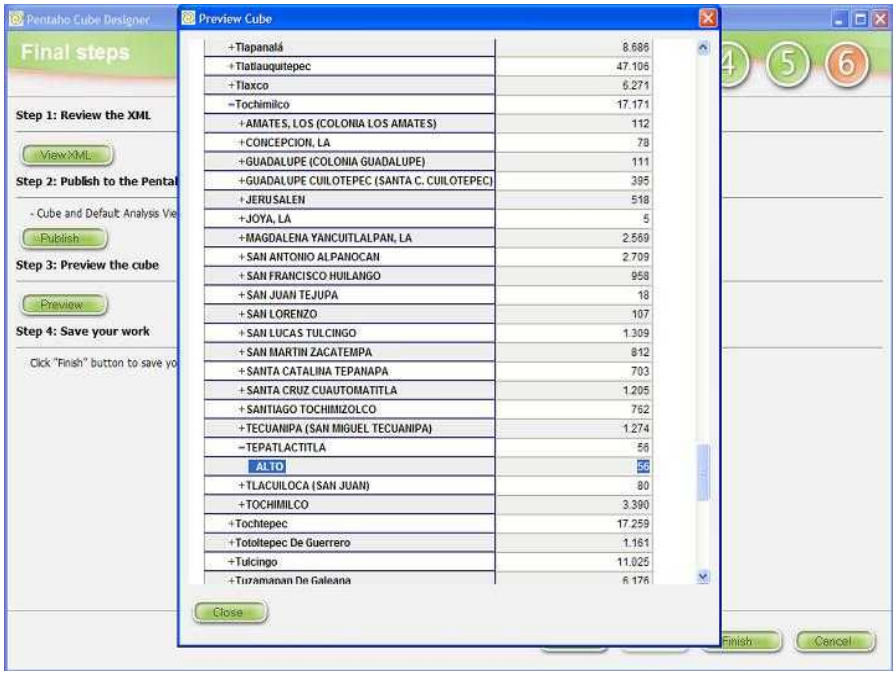


Figura 3.17 Cubo de Datos Piroclastico

En la figura 3.17 se muestra una parte del cubo de datos generado a partir de el esquema definido en la figura 3.16, aquí nos muestra la interfaz gráfica desde Cube Designer, como podemos observar, esta es la misma interfaz que si nosotros estamos en cualquier navegador web e introducimos la dirección donde se encuentra nuestro servidor Mondrian, entonces desde esta instancia podemos manipular el cubo, recordemos que todo esto se logra por medio de las expresiones multidimensionales que MDX usa para comunicarse con el Esquema generado, en la figura 3.18 se muestra la consulta en lenguaje MDX que hace posible visualizar el cubo como lo muestra la figura 3.17.

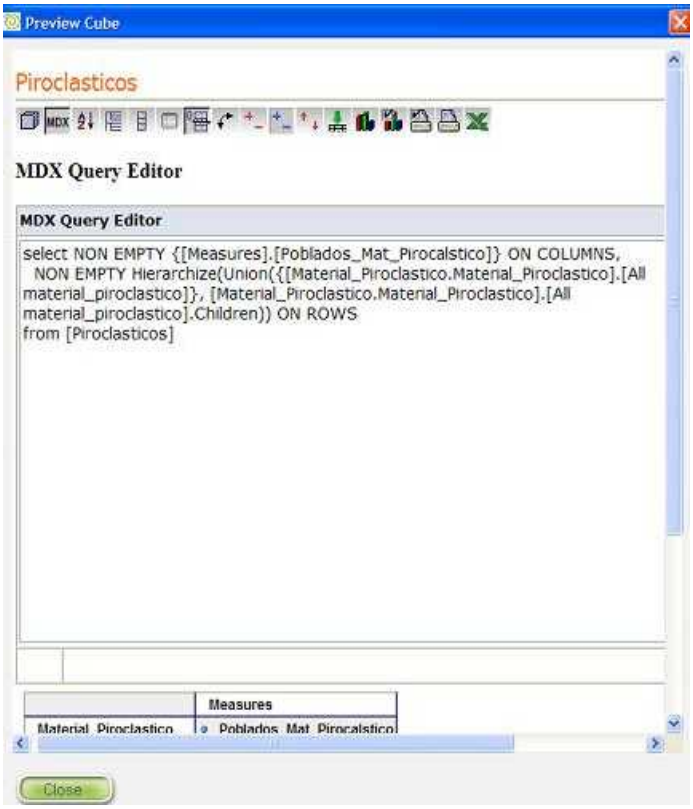


Figura 3.18 Consulta en MDX

En la interfaz de Cube Designer se nos presenta el mismo conjunto de herramientas que en Mondrian, la diferencia es que una vez que la interfaz gráfica de Cube Designer nos permite elegir que atributos serán jerarquía y cuales serán medidas, y como se van a representar dichas medidas en el cubo, solo se debe de insertar la información de manera correcta para que esta sea visualizada a manera de un cubo, ya no es necesario escribir la consulta en MDX, aunque claro, también nos permite escribir neuvas consultas con el conjunto de datos que se han tomado en cuenta en el esquema XML.

CAPÍTULO 4

CASO DE ESTUDIO

En este capítulo se hará la reseña del trabajo elaborado que sirvió como apoyo en nuestro estudio referente a la integración de los Almacenes de Datos con las Bases de Datos Espaciales, la Lógica Difusa y la Representación Multidimensional, como lo es el Cubo de Datos, a partir de las diversas herramientas que han apoyado la representación de los mismos.

4.1 Descripción del Caso

Para representar un almacén de datos espacial difuso, se propone, en este trabajo de tesis, representar las áreas de riesgo que están aledañas al volcán Popocatépetl en el estado de Puebla, mediante el análisis de mapas proporcionados por el INEGI, los cuales tienen información georeferenciada y que son de utilidad en la recopilación de la información, el fin de este estudio es tener una representación diferente haciendo uso de la manipulación de los mapas mediante diversas herramientas y clasificando esta información en conjuntos difusos, todo ello con el fin de obtener conclusiones que puedan servir en caso que se presente un acontecimiento de tipo volcánico y se necesiten tomar medidas al respecto, es decir realizar un Plan de Contingencia.

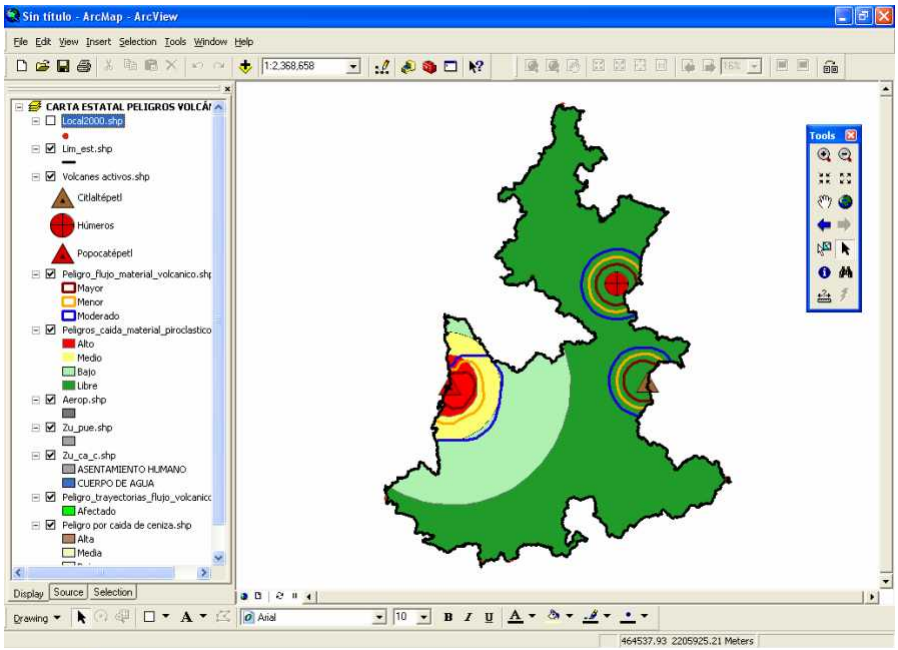


Figura 4.1 Carta Estatal de Peligros Volcánicos

En la figura 4.1 se observa uno de los mapas proporcionados por el INEGI visualizado desde el GIS ArcMap, del cual se parte para realizar las diferentes operaciones relacionadas con los objetos espaciales que se encuentran georeferenciados en las distintas capas que conforman esta representación. Nuestro caso de estudio se enfocará a la región de las tres semicircunferencias del mapa de la figura 4.1 (zonas aledañas al volcán Popocatépetl) y en algunas casos se tomará en cuenta todo el estado de Puebla (caída de material piroclástico y caída de ceniza)

Partiendo del mapa de la figura 4.1, el interés está enfocado a los eventos que pueden afectar a diversas entidades (personas, clima, regiones y geografía) en determinado momento que pudiera ocurrir un evento volcánico. Por ello se deben de tomar en cuenta las clases de objetos que pueden verse afectados como lo son las localidades en donde habitan los pobladores que viven cerca del volcán, las vías de comunicación que están disponibles cerca del lugar, el tipo de vegetación que ahí existe, el tipo de suelo que se tiene, los mantos acuíferos que no solo surten a las comunidades aledañas, sino a parte del estado de Puebla, y los distintos tipo de terrenos que pueden transportar todo el material que pueda emerger del volcán hacia las faldas del mismo (flujos volcánicos). A continuación se muestran los diferentes mapas de los cuales se tiene información relacionada a lo antes comentado.

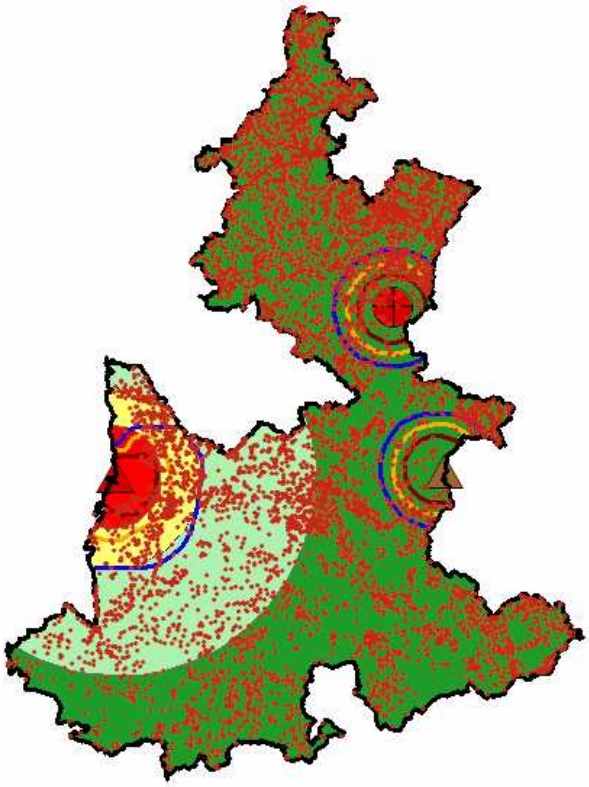


Figura 4.2 Localidades en el Estado de Puebla en el mapa de peligros

En la figura 4.2 se muestran las diversas localidades que contiene el estado. Como se puede observar las zonas que han sido clasificadas como de alto riesgo en el estado, contienen muchas localidades, mediante el uso de la herramienta que contiene ArcMap hacia las consultas espaciales que se pueden realizar se filtrarán cada una de las localidades que intervienen con cada una de las distintas zonas que contiene el mapa.

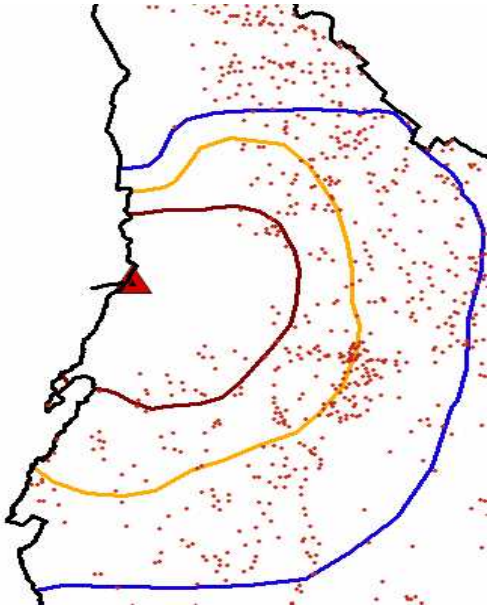


Figura 4.3 Zonas de Riesgo clasificadas por CENAPRED

En la figura 4.3 se muestra las zonas aledañas al volcán en el estado de Puebla, (haciendo un acercamiento desde ArcMap al mapa de la figura 4.1 y quitando algunas capas) las cuales han sido clasificadas como las zonas de gran riesgo por el hecho del peligro que existe en el flujo de la caída de material volcánico, como se aprecia, existen tres diferentes clases de riesgo que CENAPRED ha clasificado como ALTO, MEDIO, BAJO. Sin embargo, estas 3 clases son las que en un evento volcánico se tomarían en cuenta para llevar a cabo las medidas de protección civil, ya que como se muestra existen varios poblados o localidades que serían afectadas.

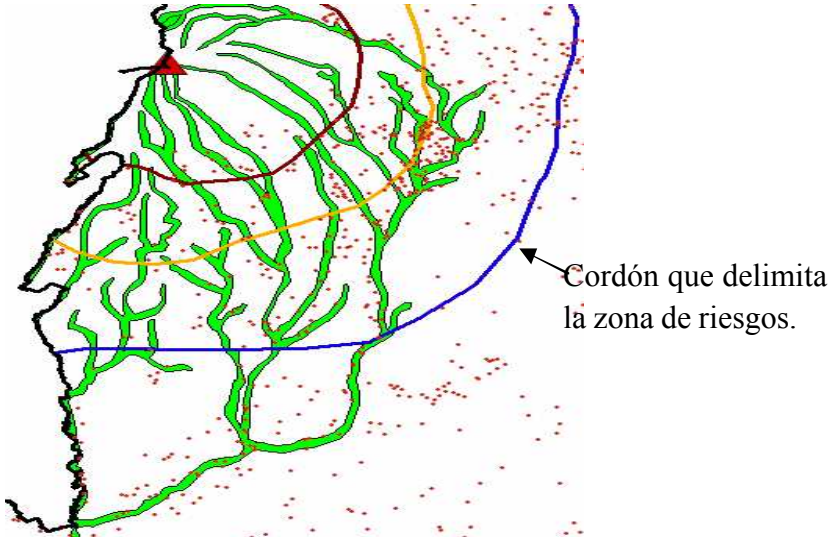


Figura 4.4 Peligro de Trayectorias del Flujo Volcánico

En la figura 4.4 y de acuerdo a la clasificación que hace INEGI, estos serían los lugares por donde correrían los flujos volcánicos, no sólo afecta al cordón q se ha clasificado como de alto riesgo sino que este sale hacia otras comunidades que quedan fuera de la zona, e incluso pasan por ellas como se puede apreciar en la figura 4.4 (parte inferior izquierda).

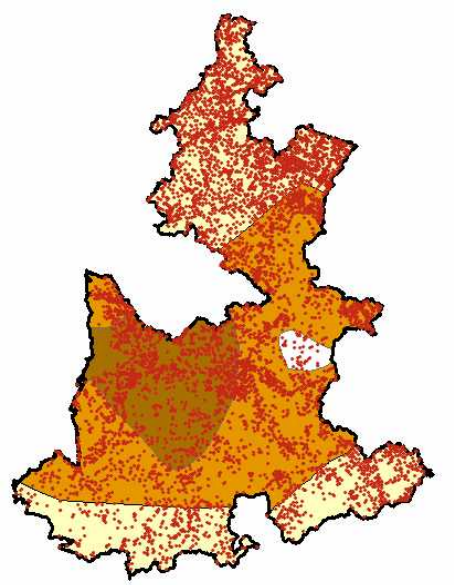


Figura 4.5 Caída de Ceniza en el estado de Puebla

En la figura 4.5 se muestra el peligro relacionado al hecho de la caída de la ceniza en el Estado de Puebla, esta clasificación también se divide en tres partes a saber Alta, Media y Baja. En este caso se debe tomar en cuenta todo el estado de Puebla, ya que la ceniza al ser un material que se desplaza por medio del viento a grandes velocidades y cubriendo grandes terrenos, provoca que el estado de Puebla, en su mayoría, se vea afectado. Incluso en el municipio de Puebla, se ha considerado que en un evento volcánico de cualquier magnitud, las casas se verían afectadas por el derrumbe de los techos por la acumulación de este material.

Otro punto que está referido a este caso de estudio es el comportamiento que ha tenido el volcán a través del tiempo, para ello se solicitó información a la BUAP CUPREDER, la cual ha estudiado el proceder del volcán con el paso de los años. Con esta información se alimentará al almacén con información referente al tiempo, esta información es el motor del almacén, ya que sin ella, no se podría realizar el análisis con el que se trabaja dentro de estas estructuras de datos.

Cabe mencionar que el estudio elaborado por la BUAP CUPREDER, abarca sólo algunos periodos en los cuales el volcán entró en una etapa de gran actividad interna, culminando con una ligera explosión, así como, deformación del cráter para después pasar a un periodo inestable por el comportamiento mostrado en el periodo que comprende de diciembre 2000 a agosto del 2003. A continuación se muestra una gráfica del comportamiento del volcán en este periodo en lo referente a exhalaciones

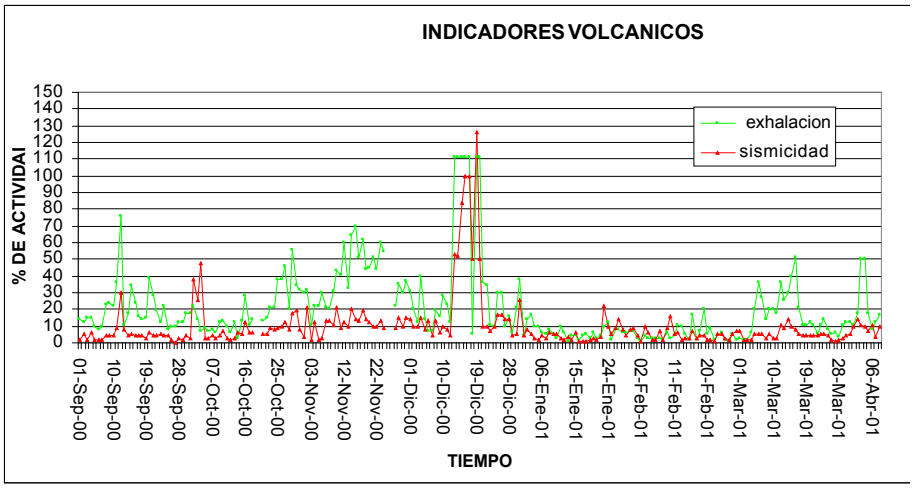


Figura 4.6 Relación de Exhalaciones del volcán en el periodo enero 2002, agosto 2003

En la figura 4.6 se observa el incremento de la actividad en INVIERNO, regularmente se ha observado un incremento de actividad volcánica, en las fechas que están próximas a invierno desde Diciembre de 1994. En el caso de la figura 4.6, se trata del año de más actividad del volcán a partir de esa fecha.

Como la clasificación de las zonas de riesgo ya está elaborada por los especialistas del INEGI y del CENAPRED, la tarea a desempeñar será el agrupar aquellas localidades que se ven afectadas en el estado de Puebla a diferentes niveles, a saber Alto, Mediano y Bajo, por los diferentes tipos de peligro que se relacionan a un evento volcánico como puede ser la caída del material piro clástico, el derrame de flujos (lava, lodo, agua) que puedan afectar a las comunidades y a sus habitantes, la caída de ceniza (en gran parte del territorio de Puebla), todo esto dependiendo del comportamiento del volcán y tomando en cuenta ciertos patrones que se han presentado antes y después de un evento volcánico.

4.1.1 Generación de las consultas espaciales

Haciendo uso de la herramienta que permite hacer selecciones de los objetos espaciales que se encuentran en una capa, que contiene ArcMap, serán separados los poblados o localidades que se ven involucradas en ello.

Por ejemplo si se desea filtrar la información referente a los nombres de los municipios que están relacionados con cada uno de los diferentes tipos de riesgos según la clasificación del CENAPRED, se deben de elegir las capas que van a participar en la selección de tales características. Por ello, únicamente se deben de tener activadas dichas capas para que las demás no participen en el momento en que se genera la selección de las características y permitan visualizar el resultado de manera correcta.

Existen dos capas que nos van a ayudar a representar este ejemplo, la primera consta de todas las localidades existentes en el estado de Puebla, representadas en la figura 4.7 por los puntos azules, la otra capa, consta de las áreas o zonas clasificadas como de gran riesgo, separadas en tres clases, ALTO, MEDIO y BAJO.

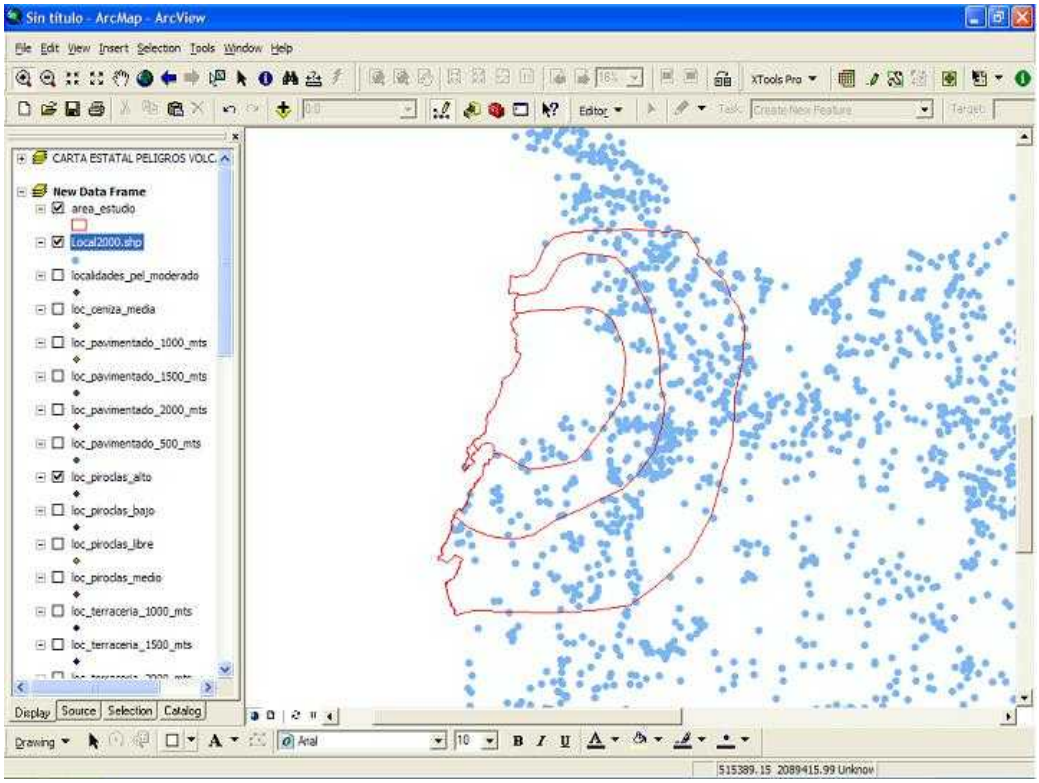


Figura 4.7 Capa de Riesgos y Localidades seleccionadas

Una vez que se han seleccionado las dos capas, se procede a realizar la selección de aquellos puntos (localidades) que están involucradas con la zona de riesgo mediante una operación espacial (within), para ello se elige de la barra de menú la opción *Selection*, luego *Select By Location* y nos desplegará una pantalla como la de la figura 4.8.

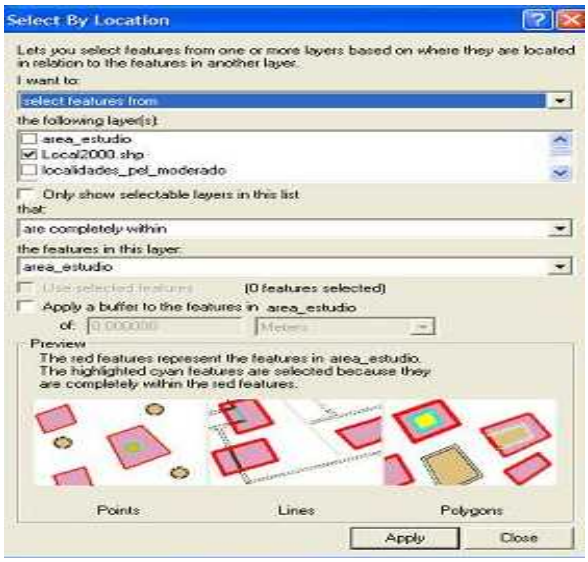


Figura 4.8 Selección de las localidades

En la figura 4.8 se muestra la manera de hacer la selección de las capas que participan en la consulta espacial, para que se seleccionen únicamente las localidades que están dentro de la zona de riesgo, una vez que se han elegido las capas a participar y el tipo de consulta que va a elaborar ArcMap se procede a ejecutar dicha operación.

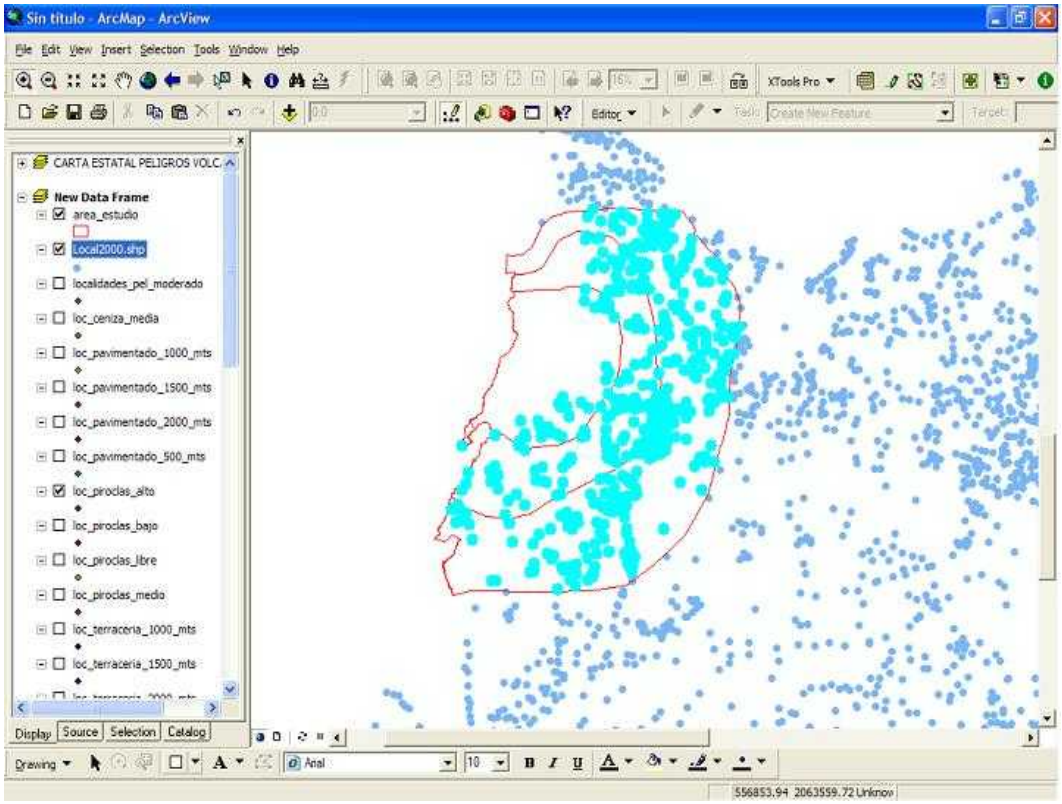


Figura 4.9 Localidades seleccionadas

Como lo muestra la figura 4.9, únicamente selecciona aquellas localidades que están adentro de la zona de riesgo, está información a su vez se selecciona de los datos almacenados en tablas de ArcMap. De esta forma podemos generar una nueva tabla que almacene estos poblados para en un futuro asociar esta capa, así como sus características a otras capas que ven involucrados a los poblados que están aledaños al volcán.

De esta misma manera podemos elegir aquellas características en las que se esté interesado. Como lo menciona el capítulo dos en la elaboración del Almacén, la información espacial con la que se cuenta y la herramienta pueden servir en esta tarea. Finalmente la figura 4.10 muestra en su mayoría, el conjunto de selecciones que se tomaron en cuenta de acuerdo a dos factores: el primero la descripción que hace CENAPRED en su sitio web acerca de los peligros asociados a el volcán y el segundo la información proporcionada por INEGI a través de sus mapas georeferenciados.

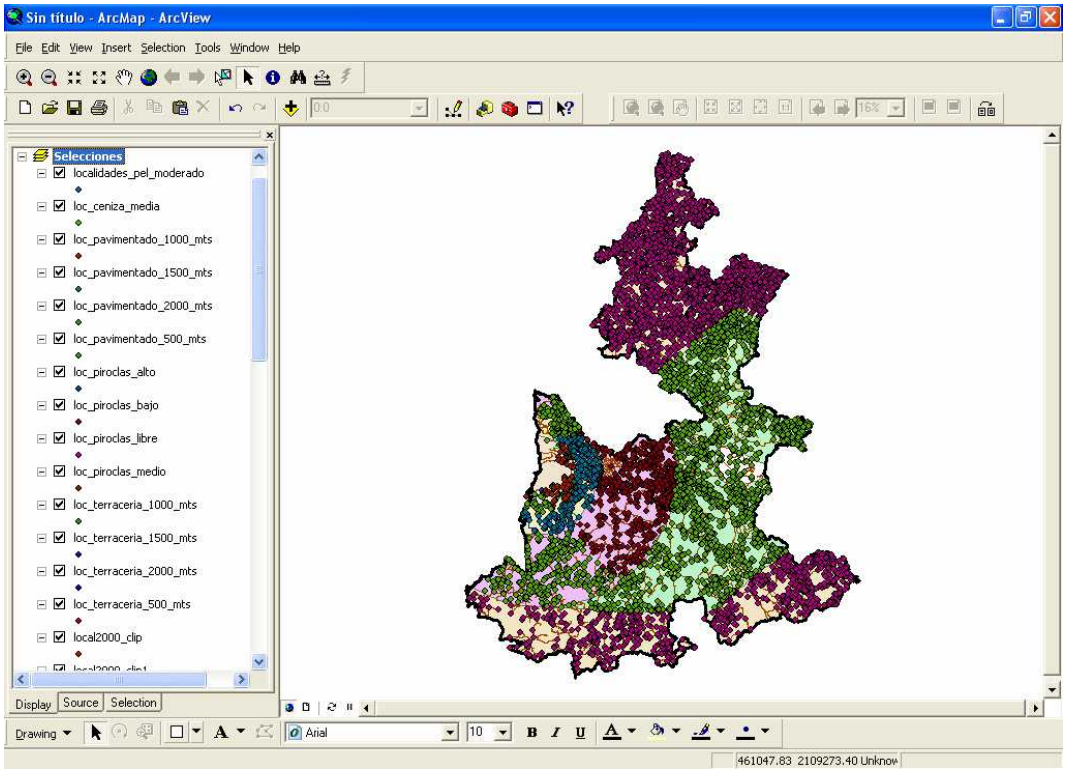


Figura 4.10 Características seleccionadas y separadas de acuerdo a criterios

En la figura 4.10 se muestran las diferentes consultas o selecciones que se elaboraron para la formación de los datos espaciales, aprovechando la herramienta que nos permite hacer selecciones de características espaciales. Esta información pasará a una base de datos en MySQL, la cual posteriormente será manipulada para generar el etiquetado de los atributos difusos (variable lingüística y grado de pertenencia).

Con ello se evita la generación de métodos de indexamiento hacia el almacén o cualquier tipo de extensión al lenguaje SQL para la formación de lo que corresponde a la dimensión espacial, y los datos quedan listos para su manejo en el almacén de datos.

En el caso de la información que está relacionada a la dimensión del tiempo, se tuvo que hacer un pequeño script que tomara la información en las tablas de Excel (formato en el que fue entregada la información por parte de BUAP CUPREDER) y pasarla a tablas de la base de datos que sirven como repositorio de datos interno. Después se le aplica otro script para asignarle la etiqueta con la que va a ser clasificado cada uno de los registros que contiene, así como lo s grados de pertenencia de esa etiqueta.

4.2 Manejo del Cubo

Supongamos que se requieren analizar los caminos que se encuentran cercanos a las comunidades que se verían afectadas por un evento volcánico importante, para ello tomamos en cuenta la información temporal con la que se cuenta, de acuerdo a la fecha en que más actividad se presentó en el periodo del 2002 a 2003. Por otra parte, tenemos que tomar en cuenta el número de pobladores que habitan ahí, por lo tanto debemos de considerar la dimensión espacial que nos ayude a recuperar todas aquellas localidades que se encuentran en mayor peligro, y que la temporada de actividad volcánica sea en invierno (fecha en que la actividad del volcán aumenta con respecto de las otras).

Estamos interesados en calcular el número de habitantes para saber en un determinado momento cómo proceder y que tipo de caminos están disponibles para poder evacuar en un plan de contingencia que se pudiera aplicar.

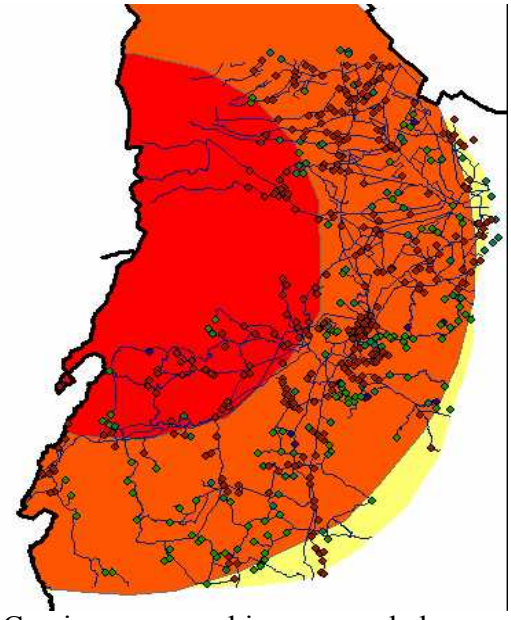


Figura 4.11 Caminos que se ubican cerca de las zonas de riesgo

En la figura 4.11 se muestran los caminos o vías de comunicación (líneas) con los que se cuenta en las localidades aledañas (puntos) al volcán, Cabe mencionar que los tipos de camino que existen en esos lugares son pavimentados y en rehabilitación, los caminos no son totalmente pavimentados, existen tramos donde se tienen caminos en rehabilitación y pedazos de camino pavimentado, por lo tanto se deben de tomar en cuenta estos tipos de camino, y como también se puede apreciar en la figura 4.11, es el hecho de que la mayoría de las localidades se encuentran en una zona de riesgo moderada, sin embargo no deja de ser peligroso para las personas que habitan ahí.

Para ello se debe de generar una nueva estructura del cubo mediante la herramienta Cube Designer, la cual pueda representar la información requerida (cantidad de pobladores afectados que puedan ser evacuados). Únicamente se tomaron en cuenta aquellos poblados que se localizan cerca (500 metros) de una vía de comunicación y que se encuentran en la zona donde el riesgo el alto.

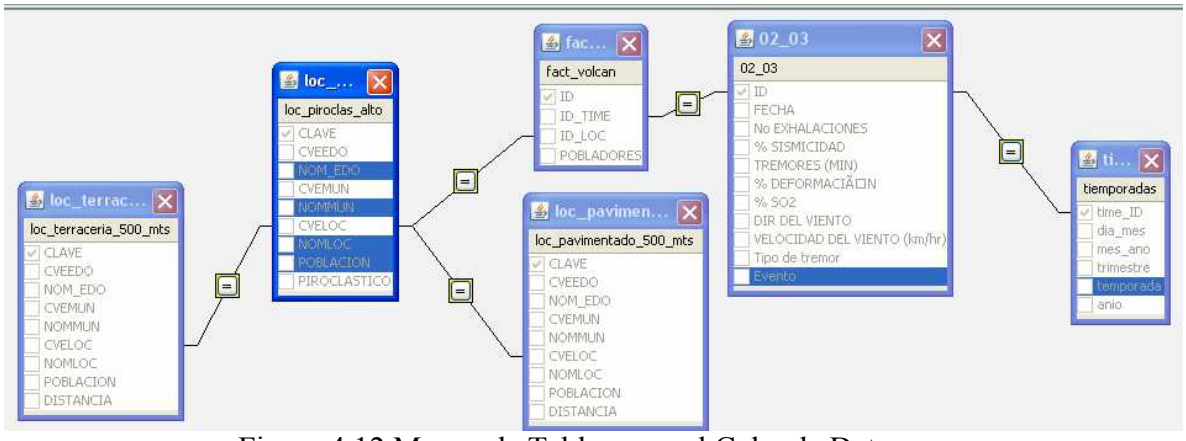


Figura 4.12 Mapeo de Tablas para el Cubo de Datos

Cabe señalar que la información ya ha sido procesada y no se ha tenido que volver a realizar la consulta desde el mapa de ArcMap para obtener los poblados que se localizan en las áreas de peligro, o el número de poblados que se encuentran a 500 o menos metros de distancia con las vías de comunicación con las que cuentan las localidades aledañas al volcán. Finalmente se muestra el resultado que arroja el cubo de datos, con cada uno de los poblados y la cantidad de pobladores afectados.

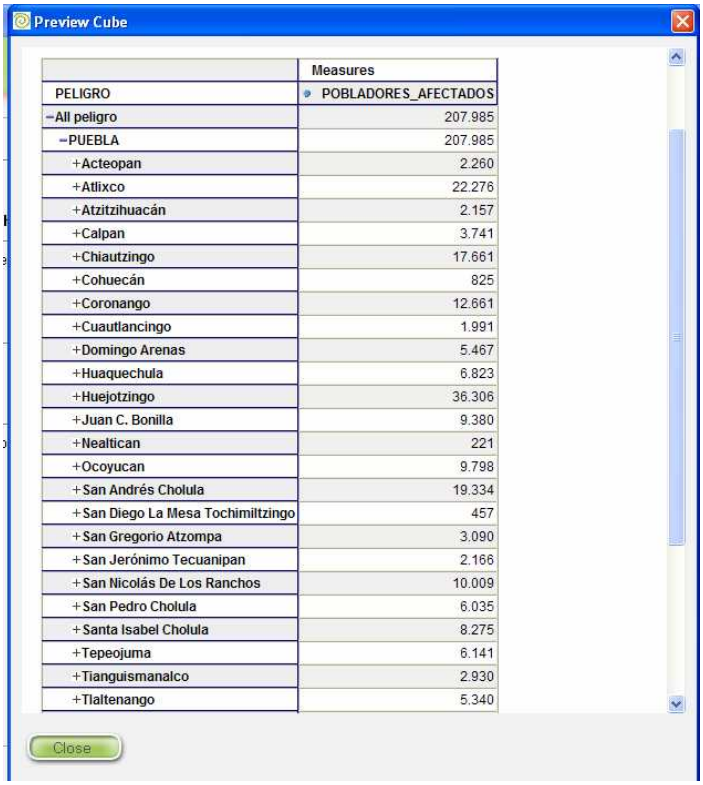


Figura 4.13 Resultados obtenidos en el cubo de datos

La cantidad de pobladores que se verían afectados por un evento volcánico, que viven en la zona de riesgo alto, y además, que las localidades se ubiquen cerca de una vía de comunicación ya sea pavimentada o en rehabilitación, suman 207,985 personas, tal y como lo muestra la figura 4.13.

CONCLUSIONES Y PERSPECTIVAS

Este trabajo representa una parte del Proyecto Geográfico Inteligente, que modelará un área geográfica de la misma manera en que aparece en el mundo real, tomando ventajas de las Tecnologías de la Información. Hemos integrado la tecnología ArcGIS con la Teoría Difusa y la representación multidimensional de la información. Una de las ventajas que tiene el trabajar con las variables lingüísticas es que están más cerca al lenguaje coloquial con el cual los seres humanos hacemos uso a diario y esto permite describir una situación geográfica de una manera más natural.

La principal contribución de este trabajo es la integración de la Lógica Difusa con las Bases de Datos Espaciales en el proceso de toma de decisiones y los procesos de consultas OLAP.

ArcGIS permite obtener características espaciales a partir de las consultas que se hacen a las capas que contienen sus mapas, haciendo una selección por características y al mismo tiempo relacionándolo con los datos no espaciales en las tablas que contienen dichos mapas. Estas características espaciales son integradas dentro una base de datos multidimensional, lo que se tiene ahora es información referente a las características espaciales, esta información ya se puede organizar en niveles (jerarquías) y permitir las operaciones de agregación y desagregación OLAP, para su representación multidimensional en un futuro. Con ello se evita la tarea al Almacén de Datos de generar los métodos de indexamiento clásicos para poder manejar los datos espaciales.

Otro aspecto es el hecho de haber agregado no sólo el valor espacial a los atributos sino darle un valor semántico (etiquetado de los atributos) al mismo, así como a otros atributos no espaciales dentro de las tres dimensiones (espacial, temática y temporal). Con esto se mejora el proceso de toma de decisiones, ya que la forma de presentar los resultados es más entendible al no presentar sólo números, sino también valores lingüísticos y se pueden generar consultas más parecidas al lenguaje coloquial que involucren a estos valores como parte de la consulta.

Finalmente, la metodología del diseño del Almacén de Datos Espaciales Difuso propuesta, simplifica el uso de las herramientas existentes para explotar el potencial del Almacén de Datos.

Como ya se ha comentado, uno de los problemas que se tuvo en la elaboración de esta investigación, en lo que se refiere al caso de estudio propuesto, fue la extracción de la información referente a la dimensión del tiempo, ya que, sólo se ha podido extraer información relacionada a un periodo de tiempo corto comparado con el necesario para poder elaborar un trabajo con buenas bases, así como la poca o nula información del tipo de vegetación, suelo, hidrología y temperatura en las zonas alrededor del volcán.

Faltaría anexar las ideas que en [6][18][23][24][25][26] se proponen de acuerdo a los métodos de indexamiento espacial existentes y lo que en [13] se proponen como modelado e integración de los conjuntos difusos con la regiones vagas al modelo presentado. Otro punto que se puede anexar a esta investigación es el trabajar con otro tipo de herramienta, (ORACLE, SQL Server) para comparar desempeños con la herramienta de Open Source (Mondrian, Cube Designer), así como las facilidades que puedan ofrecer al usuario en cuanto interfaz e instalación se refieren.

BIBLIOGRAFÍA

[1] P.A. Burrough., A. U. Frank. "Concepts and paradigms in spatial information: Are current geographic information systems truly generic?". International Journal of Geographical Information Systems, 1995.

[2] I.N. Pinto. "Técnicas de Indexado para Regiones Vagas utilizando Grid File". Tesis de Maestría. Benemérita Universidad Autónoma de Puebla. 2005.

[3] Kingston Centre for GIS, Kingston University, Kingston upon Thames, KT1 2EE, UK. Introduction to GIS and Geospatial Data, 2001.

[4] S. Shekhar, S. Chawla. Spatial Databases : A Tour. Pearson Education. Inc., 2003.

[5] S. Shekar, et al. Data Models in Geographic Information Systems. Communications of the ACM, Vol. 40 No. 4, 1997.

[6] H. Ahn, N. Mamoulis, H. Wong. A Survey on Multidimensional Access Methods. UU-CS, 2001.

[8] M. Erwing & M. Schneider. Vague Regions. 5th Int. Symp. On Advances in Spatial Databases, LNCS 1262, 298-320, 1997.

[9] J. Galindo. Conjuntos y Sistemas Difusos (Lógica Difusa y Aplicaciones). Departamento de Lenguajes y Ciencias de la Computación Universidad de Málaga, 2005.

[10] M. Fisher. Boolean and Fuzzy Regions. Department of Geography, University of Leicester, Leicester, UK., pp. 87 - 94, 2001.

[11] J. Yen and R. Langari. Fuzzy Logic, Intelligence, Control and Information Center for Fuzzy Logic, Robotics, and Intelligent Systems. Texas A & M University. Prentice-Hall, Inc., 1999.

[12] L.A. Zadeh. Fuzzy Sets. Information and Control. Vol 8, pp. 338-353, 1965.

[13] M.J. Somodevilla. Fuzzy MBRs Modeling for Reasoning about Vague Regions. Doctoral Tesis. Tulane University, 2003.

[14] Y. Wang, H. Shao. Data warehouse technology in process industry. in Intelligent Control and Automation. Proceedings of the 3rd World Congress, 2000.

[15] J. Hernández, J. R. Quintana, C. Ferri. Introducción a la Minería de Datos. Capítulos 1, 2 y 9, Pearson Prentice Hall, 2005.

[16] M. Levene, G. Loizou. Why is the Snowflake Schema a Good Data Warehouse Design? in Source, Information Systems Vol. 28 (Issue 3). pp. 225-240. ISSN 0306-4379, 2003.

[17] R. Alhajj, M. Kaya, Integrating Fuzziness into OLAP for Multidimensional Fuzzy Association Rules Mining, Third IEEE International Conference on Data Mining (ICDM'03) p. 469, 2003.

[18] V. Kamp et. al., A Spatial Data Cube Concept to Support Data Analysis in Environmental Epidemiology . 9th International Conference on Scientific and Statistical Database Management (SSDBM '97), p. 100, 1997.

[19] W.H. Inmon. Building the Data Warehouse. QED Press/John Wiley, 1992. Last edition: 3rd edition, John Wiley & Sons, 2002.

[20] S. Luján. A UML profile for multidimensional modeling in data warehouses. Data & Knowledge Engineering (DKE), 59(3), p. 725-769. ISSN: 0169-023X, 2006.

[21] M. Boehnlein, A. U. Ende .Deriving initial data warehouse structures from the conceptual data models of the underlying operational information systems. Proceedings of the 2nd ACM international workshop on Data warehousing and OLAP. p. 15-21, ISBN:1-58113-220-4, 1999.

[22] Z. Covacheva. Data warehouse architecture on the base of dimensional modelling. in International Conference on Computer Systems and Technologies - CompSysTech, p. 113-118, ISBN:954-9641-33-3, 2003.

[23] N. Stefanovic, J. Han y K. Koperski. Object-Based Selective Materialization for Efficient Implementation of Spatial Data Cubes. in IEEE Transactions on Knowledge and Data Engineering, Vol. 12, No. 6, p. 938-958, 2000.

[24] D. Papadias et. al. Efficient OLAP Operations in Spatial Data Warehouses. in Lecture Notes in Computer Science, vol. 2121. p. 443 – 459, 2001.

[25] Y. Li, Y. Chen y F. Rao. The Approach for Data Warehouse to Answering Spatial OLAP Queries. in Intelligent Data Engineering and Automated Learning. p. 270-277. SpringerLink Date Tuesday, August 26, ISBN 978-3-540-40550-4, 2003.

[26] F. Rao et. al. Spatial hierarchy and OLAP-favored search in spatial data warehouse. A survey on Multidimensional Access Methods. p. 48 - 55, ISBN:1-58113-727-3, 2003.

[30] L. Yubao, Y. Jian, "The Computation of Semantic Data Cube", GCC 2005, pp. 573-578, 2005.

[31] L. Feng, T.S. Dillon, "Using Fuzzy Linguistic Representations to Provide Explanatory Semantics for Data Warehouses", IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No 1, pp. 86-102, 2003.

Referencias de Internet

[7] http://www.udistrital.edu.co/comunidad/profesores/rfranco/calidad.htm

[27] http://mondrian.pentaho.org/documentation/architecture.php

[28] http://internap.dl.sourceforge.net/sourceforge/mondrian/Pentaho_Cube_Designer_User_Guide_0.7.0.pdf

[29] http://tomcat.apache.org/