UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

METODOS DE PARTICIONAMIENTO APLICADOS A UN ALMACEN DE DATOS DEL INEGI

T E S I S

QUE PARA OBTENER EL TITULO DE:

A C T U A R I O

P R E S E N T A :

ALBERTO GUZMAN ZEPEDA

FACULTAD DE CIENCIAS
UNAM

DIRECTORA DE TESIS: DRA. AMPARO LOPEZ GAONA

2005

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

" Métodos de Particionamiento Aplicados a un almacén de Datos del INEGI ".

realizado por Alberto Guzmán Zepeda

con número de cuenta 08507841-7 **, quien cubrió los créditos de la carrera de:** Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Dra. Amparo López Gaona
Propietario	Mat. Salvador López Mendoza
Propietario	M. en C. Juan Jesús García Gutiérrez
Suplente	Lic. en C.C. Egar Arturo García Cárdenas
Suplente	Dra. Sofía Natalia Galicia Haro

Consejo Departamental de Matemáticas

Act. Jaime Vázquez Alamilla

FACULTAD DE CIENCIAS
CONSEJO DEPARTAMENTAL
DE
MATEMATICAS

AGRADECIMIENTOS

SEÑOR, *sé que estas ahí y que mis pasos van hacia ti ... gracias por tantas bendiciones*, **TE AMO**.

Agradezco principalmente a todas aquellas personas que silenciosamente me han ayudado a alcanzar la plenitud y felicidad que hoy vivo.

Un gran respeto y admiración a todos los maestros y profesores de éste país, en especial a los que me brindaron su apoyo en la realización y revisión de este trabajo de tesis.

DOCTORA AMPARO *gracias por su disposición y amabilidad.*

MAMITA, *finalmente llegó lo que tanto estabas esperando, te lo ofrezco con mucho amor y cariño, eres la persona de la cual me siento más orgulloso; Me has enseñado a enfrentar la vida con amor y fe. Gracias por tus oraciones y por enseñarnos el camino con tu ejemplo.* **TE AMO**

PAPA, *te ofrezco este trabajo en agradecimiento a todo tu cariño y apoyo. Siempre has estado cuando te he necesitado y has sido el roble que me ha dado seguridad para emprender y enfrentar grandes retos.* **GRACIAS**

SUEGRIS, *Dios me ha consentido mucho al darme una mamita de repuesto, cuando estoy con usted, me siento en casa.* **LA QUIERO MUCHO!** (*Su yerno consentido*)

SUEGRO, *mi segundo roble, Dios me quiso bendecir al doble. Usted me ha enseñado que la vida puede verse con distintos ojos, espero aprender a verla con su sabiduría y alegría . (Gracias por su hija!!!)*

BROTHER,LIZ,ROCIO,CLAUDIA,MARTHA,ENRIQUE,ERICH,JULIO,PACOSTATION,KIKOTE,CAMANA,BETOCHAS,TATAY,ITZIAPESHOSURA,POLLITO,OMAR, ANITA, ... *son parte de mi* **TESORO**!!!

TERE,ANGIE,RAUL,???,LUISALPE..,MARIANIIITAMARIANIIITA,??? ... *le han dado un toque muy especial a mi vida en los últimos 18 años* ...**LOS QUIERO MUCHO**!

GLOBO,CHUGILO,PIKIS,HAMA,BOCASA,JULIVILLA,AARÓN,JUANJO,PEPETON, FRID,BOLIO,BOTELLO,ALEX,ALFREDO,BETO,CHAVO,CHA,COMPAFELIPE,RAFITA, ..., *que buen relajo amigos*!!! **POR LA AMISTAD**!

MALU, *gracias por tu apoyo, no dejes de buscar realizar tus sueños.*

JOS, *eres mi estrella, sigue brillando con tu entusiasmo y alegría. cuenta conmigo siempre y* **SÉ FELIZ**!!!!

MANI, *eres mi solecito, nunca dejes de poner en tu carita esa cálida sonrisa que a todos nos encanta!! con tu carisma y tenacidad lograrás lo que te propongas. Cuenta conmigo siempre y* **SÉ FELIZ**!!!!

AMOR MIO, *eres el regalo más hermoso que Dios me ha dado, tu compañía es el motor de mi vida y tu amor la luz de mi camino.* **SERÁ QUE ESTOY SOÑANDO?? TE AMO**!

TABLA DE CONTENIDO

Introducción

En la actualidad las organizaciones modernas requieren sistemas de soporte a la toma de decisiones, los cuales están basados en información proveniente de diversas fuentes. En ocasiones, dichos sistemas acumulan cientos de *gigabytes* o varios *terabytes* de información en lo que hoy en día se conoce como almacenes de datos o *datawarehouse.*

Estas organizaciones enfrentan grandes retos para el mantenimiento y buen rendimiento de sus almacenes de datos debido al gran volumen de datos que manejan, y uno de los métodos para solventar estos retos es el *particionamiento* de datos, el cual consiste en mecanismos implementados en los manejadores de bases de datos para organizar físicamente los datos de manera eficiente sin afectar la estructura lógica definida para los mismos.

Al conocer los métodos de particionamiento que una base de datos relacional puede soportar, éstos pueden ser considerados desde el diseño del almacén de datos, lo cual garantizará en etapas tempranas del diseño, un fácil mantenimiento y buen rendimiento del sistema.

El presente trabajo de tesis tiene como objetivo aplicar métodos de particionamiento en el almacén de datos del Instituto Nacional de Estadística, Geografía e Informática que contiene el censo de población y vivienda realizado en el año 2002 y explicar los beneficios obtenidos al aplicar dichos métodos.

Para lograr el objetivo planteado, este trabajo se estructura de la siguiente forma:

En el capítulo 1 se presentan conceptos sobre almacenes de datos que son utilizados en capítulos posteriores.

En el capítulo 2 se exponen métodos de particionamiento de datos comúnmente utilizados para la optimización del mantenimiento y rendimiento de almacenes de datos.

En el capítulo 3 se presenta como caso de estudio, la síntesis metodológica del censo nacional de población y vivienda realizado por el INEGI en el año 2000.

En el capítulo 4 se realiza el análisis del caso de estudio y se determinan los métodos de particionamiento adecuados para el almacén de datos que contiene la información del censo.

Finalmente en el capítulo 5 se presentan las conclusiones obtenidas de aplicar los conceptos teóricos, al caso de estudio presentado.

Capítulo 1

Conceptos generales de almacenes de datos

En este capítulo se presentan conceptos generales sobre almacenes de datos, que son utilizados en el desarrollo de los capítulos posteriores.

Los temas desarrollados en este capítulo están orientados a establecer un marco teórico que permita comprender la necesidad de implementar métodos de particionamiento en almacenes de datos.

Se define lo que se conoce como un almacén de datos en el ámbito informático y sus características peculiares que lo diferencian de otros sistemas informáticos. También se presentan las arquitecturas tecnológicas que comúnmente se utilizan para implantar un almacén de datos y los aspectos a considerar en el diseño físico y lógico del mismo.

Los temas a desarrollar en este capítulo son:

- Qué es un almacén de datos
- Características de un almacén de datos
- Diferencias entre un almacén de datos y un sistema transaccional
- Arquitectura de un almacén de datos
- Diseño lógico y diseño físico

1.1 Qué es un almacén de datos

Un almacén de datos es una base de datos que concentra e integra datos provenientes de diversas fuentes, los cuáles, son acomodados especialmente para ser analizados y con base en este análisis, tomar decisiones inteligentes. [1]

Las características y diseño de un almacén de datos difieren de las bases de datos que soportan los sistemas transaccionales tradicionales, entendiendo por sistema transaccional aquel que se utiliza para registrar y controlar la operación diaria de una organización.

Los usuarios de un almacén de datos, utilizan la información contenida en éste para analizar de manera oportuna el comportamiento de una organización, dado que sin la existencia de un almacén de datos, la oportunidad para analizar la información se disminuye por el tiempo que se invierte en extraer la información para después analizarla. Es decir que el diseño y la creación de un almacén de datos se justifica para poner a disposición de los analistas los datos acomodados y listos para su explotación y comprensión y así, dediquen el 100% de su tiempo al análisis sin invertir tiempo en la obtención y acomodamiento de las fuentes de información.[2]

El fin primordial de un almacén de datos es concentrar la información de forma tal que sea posible generar hipótesis, simular decisiones,

---

[1] Data Warehousing, US, Digital Press, Lilian Hobbs p 1-14

[2] Bussiness Intelligence, Gartner Publications p-3-16

descubrir tendencias y llevar a la organización a tomar decisiones inteligentes.

De lo anterior se deriva el concepto de inteligencia de negocios o *Business Intelligence* que está íntimamente relacionado con el concepto de almacén de datos o *Datawarehouse* [3].

La inteligencia de negocios, busca ayudar en la comprensión de los datos para tomar mejores decisiones de manera oportuna y finalmente alcanzar los objetivos de una organización de manera efectiva y eficiente. Una de las metas de la inteligencia de negocios es hacer que el flujo de la información dentro de una organización sea flexible y accesible, otra es encontrar los elementos necesarios para rediseñar los procesos de la organización minimizando costos e incrementando la rentabilidad (o la efectividad en el caso de una organización gubernamental o altruista). A pesar de llamarse inteligencia de negocios, esta tendencia aplica para cualquier tipo de organización que pretenda mejorar su operación y maximizar resultados. Por ejemplo en una organización gubernamental o altruista, la toma de decisiones debe estar basada en canalizar los recursos para beneficiar a la mayor cantidad de población posible, como es el caso del INEGI, que concentra información que permite a otras entidades conocer las características y tendencias de la población para decidir la mejor forma de canalizar recursos para beneficiar a los sectores más necesitados o atender áreas que provean un mayor crecimiento económico.

Un almacén de datos se crea cuando existe la necesidad de concentrar e integrar la información que proviene de diversas fuentes y se

[3] Ralph Kimball, es reconocido como uno de los padres del concepto de Datawarehouse, se ha dedicado desde hace más de 10 años al desarrollo de su metodología para que éste concepto sea bien aplicado en las organizaciones y se asegure la calidad en el desarrollo de estos proyectos.

requiere conservar la mayor cantidad de información histórica posible. Al integrar esta información, también se busca acomodarla de tal forma que su acceso sea eficiente y útil para el análisis y toma de decisiones oportunas, para lo cuál se contemplan las siguientes etapas que cubren el ciclo de vida de un almacén de datos:

- Extracción
- Carga
- Transformación
- Explotación
- Depuración

La etapa de *extracción* consiste en obtener y poner en un formato establecido, los datos de interés provenientes de diversas fuentes. Normalmente estas fuentes son bases de datos de sistemas administrativos de la organización. Pero en ocasiones los datos pueden provenir de otras fuentes como encuestas, estadísticas, muestras, etcétera.

Durante la etapa de *carga*, los datos extraídos se transfieren a bases de datos temporales para transformarse en la siguiente etapa.

Finalmente en la etapa de *transformación* los datos depositados en las áreas temporales se acomodan de acuerdo a un modelo de datos orientado al análisis y explotación de información histórica.

Dado que los almacenes de datos suelen acumular grandes cantidades de información histórica, y los recursos de cómputo siempre son limitados, la etapa de *depuración* es esencial para eliminar los datos que tienen mayor antigüedad y liberar espacio para datos más recientes[4].

---

[4] **Oracle® Database Data Warehousing Guide**
10g Release 1 (10.1) Cap. 11 Overview of Extraction, Transformation, and Loading

Por lo anterior un sistema de soporte a la toma de decisiones basado en un almacén de datos debe contemplar herramientas que faciliten las siguientes tareas:

- Diseño lógico y físico del modelo de datos orientado al análisis y explotación de información.
- Diseño, automatización y control de procesos de extracción, transformación y carga.
- Explotación de la información mediante la elaboración de reportes predefinidos o informes definidos por el usuario de manera ocasional o arbitraria.
- Procesamiento Analítico en Línea (OLAP) con capacidades de estimación, proyección etc.

1.2 Características de un almacén de datos

Según la definición inicial de un almacén de datos, que establece que concentra e integra información de diversas fuentes, se tienen como características principales:

a) Orientado a temas específicos.
b) Integrado
c) No volátil
d) Variante en el tiempo

Se detallan a continuación las características mencionadas:

1.2.1 Orientado a temas específicos

Esta característica se refiere a que un almacén de datos está diseñado para ayudar con el análisis de los datos sobre un tema específico. Por ejemplo, para tener un mayor conocimiento acerca de los datos relevantes a las ventas de una compañía, se puede crear un almacén de datos orientado a las ventas de la compañía. Usando este almacén de datos, se podrían responder preguntas tales como:

¿Quién fue el mejor cliente en algún producto en particular el año pasado?

Para efectos de la aplicación práctica de este trabajo de tesis, que está basado en un almacén de datos del INEGI, el tema específico de este almacén de datos son las características socio-demográficas de la población de los diversos estados de la republica, que servirá para responder a preguntas tales como:

¿Cuál fue la distribución de la población según los grandes grupos de edad y sexo en el año 2000?

¿Cuál fue la distribución de la población según su condición migratoria municipal, estatal e internacional en el año 2000?

¿Cuál fue la distribución de la población discapacitada, según el tipo de discapacidad, en el año 2000?

¿Cómo se distribuye la población laboralmente activa respecto a la edad?

Por tal motivo, un almacén de datos tiene como característica estar orientado a temas específicos (subject oriented)

Lo que se puede observar en las preguntas anteriores es que el tema específico que se desea analizar es la población bajo diversas dimensiones tales como el estado laboral, la edad, el sexo, estado migratorio etc.

1.2.2 Integrado

El concepto de integración está estrechamente relacionado con el concepto de orientación a un tema específico, y se refiere a que los almacenes de datos deben poner los datos provenientes de diversas fuentes de una forma consistente.

Deben resolver problemas tales como conflictos en la conceptualización de aspectos del negocio o inconsistencias en unidades de medida etc.; cuando resuelven este tipo de problemas, se dice que son almacenes de datos integrados.

Para el caso del almacén de datos del INEGI, se integra información de diversos aspectos socio-demográficos de la población como son la edad, el sexo, la condición migratoria, el ingreso per capita etc. Y dichos aspectos se deben de alinear a estándares previamente establecidos como se explica en el capítulo 4.

1.2.3 No volátil

Significa que una vez introducidos los datos en el almacén de datos, después de las etapas de extracción, transformación y carga, éstos no deben cambiar. Sólo se depuran o se incrementa su volumen.

En una base de datos que soporta un sistema transaccional, los datos cambian continuamente de acuerdo a la operación de la organización.

Por ejemplo la operación diaria de un banco implica cambios continuos a los saldos de los cuenta habientes, lo cual genera cierta "volatilidad" en los datos, pero una vez consolidados estos movimientos en un almacén de datos, éstos quedan como referencia histórica que permitirá estudiar el comportamiento de las operaciones a través del tiempo.

1.2.4 Variante en el tiempo

Para poder descubrir tendencias en el negocio, los analistas necesitan grandes cantidades de información, en contraste con los sistemas transaccionales, donde los requerimientos de rendimiento demandan que la información histórica sea depurada. Un almacén de datos se enfoca en el comportamiento de los datos a través del tiempo, por ello la característica de ser variante en el tiempo.

Típicamente los datos fluyen de un sistema transaccional a un almacén de datos en períodos mensuales, semanales o diarios. Los datos normalmente son procesados en áreas o archivos temporales antes de ser integrados al almacén de datos. Los almacenes de datos comúnmente ocupan cantidades que van desde cientos de *gigabytes* hasta varios *terabytes* de datos y usualmente esta información se concentra en tablas históricas muy grandes que contienen los hechos ocurridos en el negocio con el transcurrir del tiempo, que son de interés de análisis para analistas del negocio.

Lo anterior no contradice la característica de "no volatilidad" dado que el hecho de ser "variante en el tiempo" sugiere conservar la mayor cantidad de información histórica que no debe cambiar salvo que tenga que ser depurada por cuestiones de espacio.

1.3 Diferencias entre un almacén de datos y una base de datos de un sistema transaccional

Partiendo del concepto de que un sistema es un conjunto de elementos interactuando para producir uno o varios resultados específicos, se entiende que un sistema informático en términos generales se compone de un conjunto de unidades de programación (formas, reportes, procesos, menús etc.) y una base de datos que almacena la información que es generada o modificada por dicho conjunto de unidades de programación.

La base de datos que soporta un sistema transaccional, es diferente a la base de datos que soporta un sistema de toma de decisiones.

Los sistemas transaccionales (*On Line Transacction Procesing* OLTP) son los sistemas tradicionales diseñados para controlar los flujos diarios de información durante operación de una organización y sus bases de datos tienen requerimientos diferentes con respecto a los almacenes de datos de sistemas de soporte a la toma de decisiones (*Decission Support Systems* DSS), los cuales se mencionan a continuación:

1.3.1 Carga de trabajo

Los almacenes de datos están diseñados para atender consultas planeadas o no planeadas, es decir, consultas que se plantean de manera espontánea por los analistas o tomadores de decisiones, y posiblemente por esta razón no sea posible conocer anticipadamente la carga de trabajo que tendrá un almacén de datos a diferencia de una base de datos de un sistema transaccional que tiene bien definidas sus consultas dentro de sus procesos operativos.

Por lo anterior, un almacén de datos debe diseñarse para atender óptimamente una gran variedad de operaciones de consulta.

Los métodos de particionamiento expuestos en el capítulo 2 son una forma de atender a este requerimiento.

1.3.2 Modificaciones a los datos

Un almacén de datos se actualiza de manera regular mediante los procesos de extracción, transformación y carga previamente diseñados y automatizados.

Los usuarios finales del almacén de datos no actualizan directamente los datos a diferencia de las bases de datos de los sistemas transaccionales que reciben continuamente modificaciones de los usuarios u operadores del sistema y reflejan el estado actual de cada transacción u operación propia del negocio.

En el caso de un almacén de datos, la función primordial de los usuarios es consultar la información, analizarla, detectar comportamientos y tendencias para tomar decisiones, en el caso de un sistema transaccional los usuarios son los que hacen uso de las unidades de programación (pantallas, procesos, reportes etc.) para reflejar las operaciones que se realizan en la organización (afectaciones al inventario, afectaciones contables etc.)

1.3.3 Modelo de datos

Los modelos de datos frecuentemente usados en un almacén de datos se conocen como "modelos de estrella"[5] que están representados por una tabla que concentra la información histórica de los hechos o acontecimientos que son el sujeto de análisis, y un conjunto de tablas circundantes llamadas dimensiones que son los atributos o aspectos bajo los cuales se desean visualizar los hechos.

Un ejemplo de modelo de estrella clásico es una tabla de hechos que refleja las ventas realizadas durante los últimos 5 años y 3 dimensiones al rededor de ella: un catálogo de productos, un catálogo de regiones y un catálogo de clientes, en donde lo que se desea es analizar las ventas por producto, región y cliente.

Los modelos de estrella se explican con mayor detalle más adelante en este capítulo en la sección 1.5.1.1 del diseñó lógico de un almacén de datos.

Frecuentemente los almacenes de datos están basados en modelos de datos denormalizados o parcialmente normalizados, lo cual se refiere a que no cumplen o cumplen parcialmente con algunas reglas básicas del modelo de datos relacional, las cuáles establecen que:

1) Las tablas deben tener una llave primaria que asegure que todos y cada uno de los registros sean únicamente identificables.
2) Las columnas que forman parte de una llave primaria deben ser no nulas y su combinación debe ser única.

---

5 Oracle® Database Data Warehousing Guide
10g Release 1 (10.1) Cap. 4 Datawarehouse Models

3) Una llave foránea debe hacer referencia a la llave primaria o una llave única de la misma o de otra tabla.

A diferencia de un almacén de datos, los datos en un sistema transaccional deben cumplir al 100% las reglas de normalización de bases de datos relacionales para garantizar la consistencia de los datos y el apego a las reglas del negocio.

La razón por la cual un almacén de datos no se adhiere al 100% a las reglas de normalización, es por que el gran volumen de datos que maneja ocasionaría tiempos de respuesta no aceptables en los procesos de transformación, carga o explotación de la información.

1.3.4 Operaciones típicas

Las consultas realizadas en un almacén de datos típicamente recorren miles o millones de registros de una tabla para agruparlos, por ejemplo "encontrar las ventas totales de los últimos 5 años agrupadas por año, producto y región" es una consulta que requiere recorrer todos los registros de la tabla de hechos (ventas) y agruparlos por las dimensiones: año, producto y región.

En el caso de un sistema transaccional, las consultas están orientadas a registros específicos que son encontrados mediante la utilización de índices, por ejemplo "obtener la orden 2467 del cliente 345" requerirá utilizar un índice sobre la tabla de ordenes para localizar la orden 2467.

1.3.5 Datos históricos

Los almacenes de datos usualmente almacenan muchos meses o años de información para satisfacer las necesidades análisis histórico de la información a diferencia de los sistemas transaccionales que por razones de desempeño y conveniencia depuran la información histórica para mantener sólo algunos meses de información en línea.

1.4 Arquitectura de un almacén de datos

Se presentan a continuación tres arquitecturas básicas que se emplean en almacénes de datos:

1.4.1 Arquitectura básica

La arquitectura básica de un almacén de datos consiste en concentrar en una sola base de datos los metadatos (la descripción de la información), los datos provenientes de los sistemas transaccionales y tablas con datos resumidos.

Bajo esta arquitectura se tiene la ventaja de concentrar en un solo lugar los datos alineados modelos de análisis de información, es decir, los datos acomodados de tal forma que puedan ser consultados y analizados eficientemente; y los datos provenientes del sistema transaccional, lo cual facilita el análisis directo de la información cuando se desea acceder un mayor nivel de detalle.

Por ejemplo, al visualizar las ventas del 2003, podemos preguntarnos cuáles son las 10 órdenes de compra con mayor volumen de ventas en ese año, lo cual nos llevaría a consultar directamente bajo esta arquitectura las ordenes de compra del 2003 en el almacén de datos.

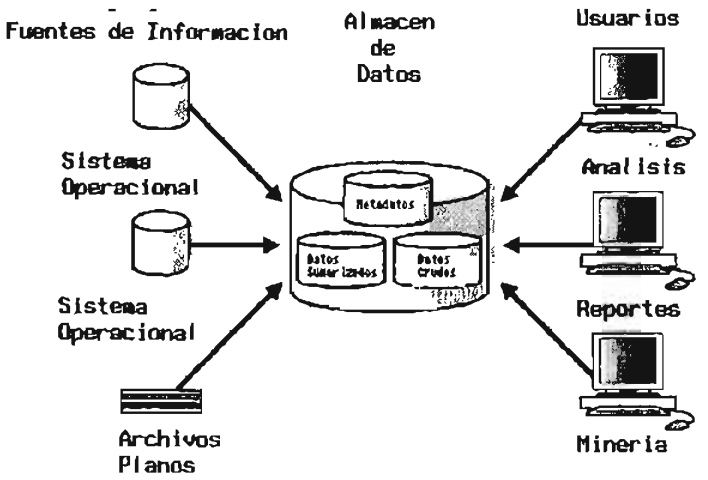


Figura 1 Arquitectura Básica de un Almacén de Datos[6]

En la figura 1 se observa la arquitectura básica donde las fuentes de información son los sistemas operacionales diversos, y/o archivos planos donde se ha depositado información de interés. Por otro lado se tienen diversas clases de usuarios, que podrían ser analistas financieros, analistas de tendencias (minería) etcétera, que consultan directamente el almacén de datos. Al centro se observa la información

[6] **Oracle® Database Data Warehousing Guide 10g Release 1 (10.1) Cap.** 1 Datawarehouse Concepts

clasificada en metadatos, los cuales describen lo que representa cada pieza de información; los datos crudos, que son los datos tal como provienen de su fuente y los datos resumidos que como su nombre lo indica, acomodan los datos crudos de forma resumida para hacer eficientes ciertas consultas típicas.

1.4.2 Arquitectura con áreas temporales de trabajo

Esta arquitectura contempla una base de datos temporal de trabajo con sus propios recursos de cómputo destinados a las etapas de carga y transformación del almacén de datos.

Cuando los procesos utilizados para transformar los datos, son muy complejos y demandan grandes cantidades de recursos de cómputo, se utilizan las áreas temporales de trabajo que se mencionan en esta arquitectura.

De ésta forma, no se interrumpen ni se utilizan los recursos de cómputo destinados para la explotación de la información y se pueden acomodar datos provenientes de otras diversas fuentes para finalmente tener un almacén de datos ajustado al modelo deseado ya sea dentro de la base de datos temporal o en otra base de datos.

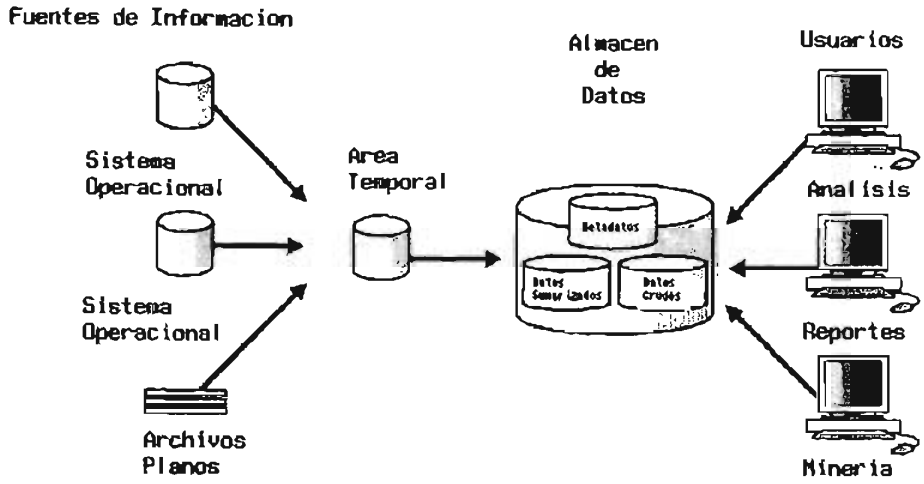


Figura 2 Arquitectura con areas temporales de trabajo[7]

1.4.3 Arquitectura con Áreas Temporales de Trabajo y Subalmacenes de Datos (Datamarts)

Tomando en cuenta la característica de que los almacenes de datos están orientados a temas específicos, se puede tener una base de datos temporal para la extracción, transformación y carga de datos que contenga la información necesaria para crear sub almacenes de datos o DATAMARTS con una orientación especifica, por ejemplo, una vez finalizadas las etapas de extracción, información y carga, se puede tener como producto final un *datamart* de ventas, un *datamart* de inventarios y un *datamart* de compras.

Los *datamarts* o sub almacenes de datos serán entonces subconjuntos de un almacén de datos global y tendrán un tema especifico. Estos

[7] **Oracle® Database Data Warehousing Guide 10g Release 1 (10.1) Cap.** 1 Datawarehouse Concepts

datamarts pueden ubicarse en la base de datos temporal de procesamiento o en otra base de datos con mayor accesibilidad a los usuarios que analizarán la información.

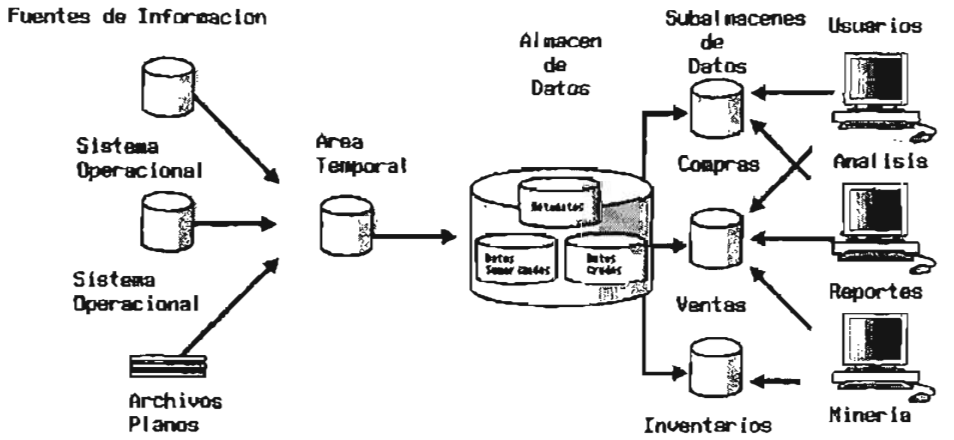


Figura 3 Arquitectura con áreas temporales y subalmacenes de datos[8]

Las tres arquitecturas básicas mencionadas anteriormente contemplan la creación de tablas que precalculan operaciones que normalmente se realizan en el análisis de la información, por cuestiones de optimización en los tiempos de respuesta de las consultas.

Por ejemplo en un almacén de datos orientado a las ventas de una compañía, normalmente los analistas realizarán consultas que agrupen la información por producto o por período, lo cual puede quedar precalculado en una tabla con estas agrupaciones y a partir de esta tabla realizar consultas de mayor complejidad, eliminando la necesidad de agrupar la información cada vez que un usuario lo requiera.

[8] **Oracle® Database Data Warehousing Guide 10g Release 1 (10.1) Cap.** 1 Datawarehouse Concepts

Uno de los aspectos importantes a considerar en la arquitectura de una almacén de datos, es el tipo de servidores de base de datos o *hardware* que se utiliza para soportar la operación diaria.

1.4.4 Arquitecturas de hardware y paralelismo

En la actualidad, tanto los sistemas transaccionales como los sistemas de soporte a la toma de decisiones, utilizan bases de datos que se encuentran en servidores con capacidades de procesamiento especiales. Estas capacidades ofrecen la posibilidad de contar con una arquitectura de multiprocesamiento, lo cual se refiere a que los servidores son escalables o pueden crecer en el número de procesadores (CPU's) que utilizan para dar atención a las peticiones del sistema operativo o el manejador de base de datos.

Con las capacidades de multiprocesamiento mencionadas, se pueden entonces "paralelizar" tareas en el servidor que tienen un alto consumo de recursos de disco y procesador, para que éstas sean eficientes. Por ejemplo, en el caso de un almacén de datos normalmente las consultas requieren recorrer grandes cantidades de registros para resolverse, lo cual se traduce en realizar muchas operaciones de acceso a disco (I/O) para obtener los datos solicitados. Cuando dicha operación de acceso al disco (I/O) no se realiza en paralelo, tenemos entonces solamente un proceso en el sistema operativo realizando ésta operación, de tal forma que si cada operación de I/O toma 1 milisegundo y para realizar la consulta se tienen que realizar 10 millones de operaciones de I/O, entonces tomaría 10 millones de milisegundos (10 mil segundos) en finalizarse; Pero si dicha consulta se realiza con paralelismo, se podrían tener varios procesos simultáneos, sobre los que se repartirían las operaciones de I/O, de tal forma que si se generan 10 procesos para

realizar los 10 millones de operaciones de I/O, entonces cada proceso realizaría 1 millón de operaciones y la consulta tardaría 1 millón de milisegundos (1000 segundos), es decir que si una operación no paralelizada toma M segundos en realizarse, al paralelizar dicha operación con N procesos, el tiempo de finalización se reduce a M/N.

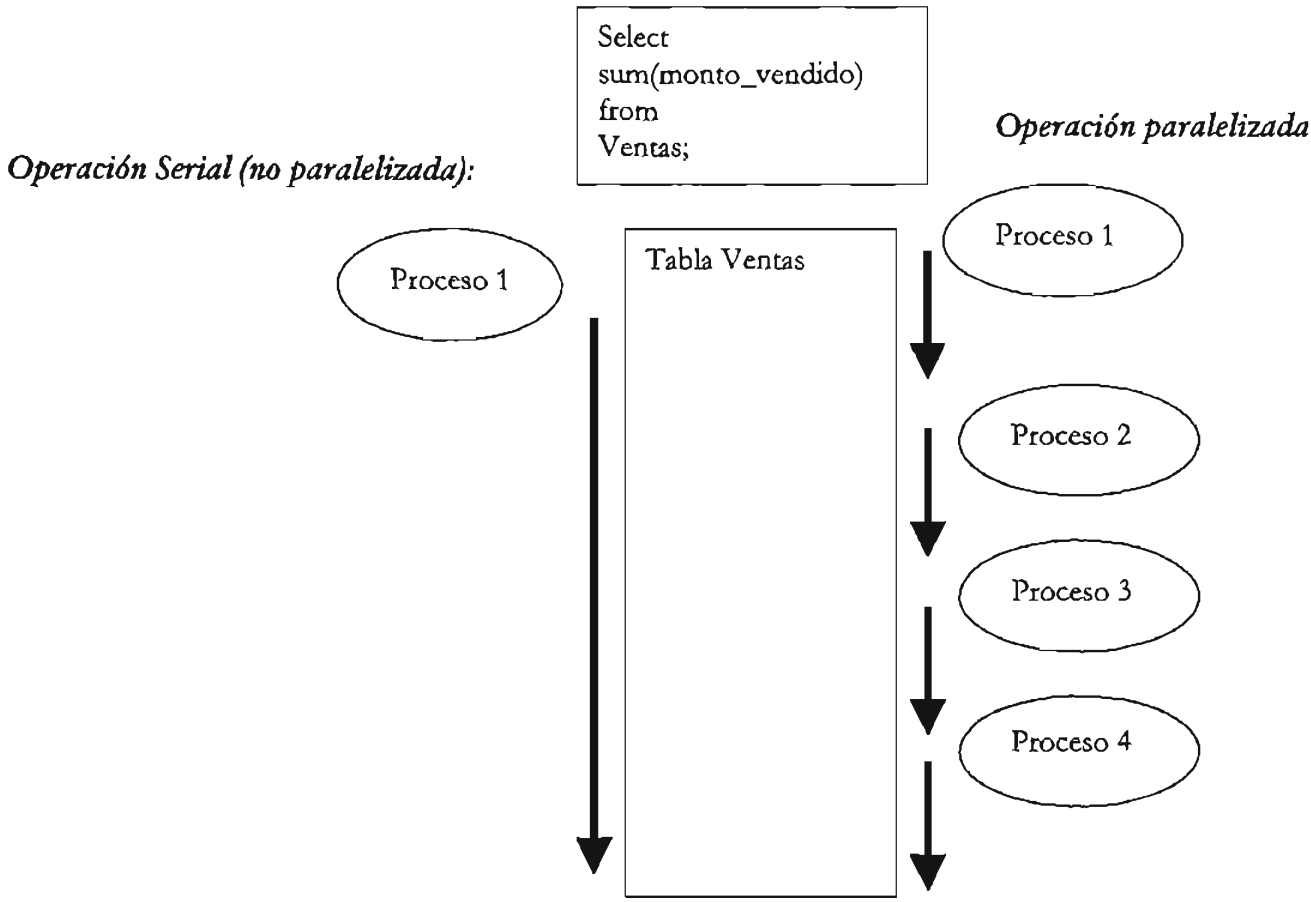


Figura 4 Comparación entre procesamiento serial y paralelo

La idea de la paralelización es lograr el mismo resultado en menos tiempo pero con más recursos de hardware, lo cual tiene un precio que no siempre puede pagarse. Por tal motivo, es necesario al momento de diseñar una arquitectura de un almacén de datos y al realizar el diseño físico del mismo, considerar que los recursos de *hardware* son limitados y que deben optimizarse al máximo, por lo que una buena combinación de paralelismo en algunas operaciones y de organización física de los datos es la clave de una buena arquitectura [9].

Es importante también considerar que si se estiman volúmenes de procesamiento altos, se puede optar por realizar planeación de capacidad de los recursos de *hardware* que se necesitarán para soportar el sistema que se desea implantar, lo cual consiste en que una vez que se tienen finalizados distintos diseños para el sistema (desarrollados primero en ambientes de *hardware* reducidos), después se realizan pruebas en equipos rentados temporalmente para determinar el *hardware* necesario, minimizando su costo mediante diversas técnicas, una de las cuales son los métodos de particionamiento de datos que se exponen en el siguiente capítulo.

1.5 Diseño de un almacén de datos

Una vez que una organización decide construir un almacén de datos y determina el alcance y uso que éste tendrá, se traducen los requerimientos en el diseño de un sistema que los cubra mediante un diseño lógico y físico que contempla:

---

[9] Oracle® Database Data Warehousing Guide 10g Release 1 (10.1) Cap. 5 **Parallelism and Partitioning in Data Warehouses**

- El contenido específico o temas de interés.
- Las relaciones entre los grupos de datos.
- El ambiente tecnológico que soportará el sistema
- Las transformaciones de datos requeridas
- La frecuencia de refrescamiento de los datos

El diseño lógico es más conceptual y abstracto que el diseño físico. En el diseño lógico, se buscan las relaciones lógicas entre los objetos. En el diseño físico, se busca la manera óptima para almacenar y extraer los datos, así como también para manejarlos, administrarlos, depurarlos y respaldarlos.

El diseño debe orientarse hacia las necesidades de los usuarios finales. Los usuarios finales típicamente realizan análisis y se enfocan hacia datos agrupados en lugar de transacciones aisladas o individuales. Sin embargo, los usuarios finales podrían no saber lo que realmente necesitan hasta que lo ven. Adicionalmente, un buen diseño planeado debe permitir el crecimiento y los cambios conforme las necesidades de los usuarios cambian.

Al principio, en el diseño lógico, el enfoque es hacia los requerimientos de información para después preocuparse por los detalles de implementación que se derivan del diseño físico.

1.5.1 Diseño lógico

Un diseño lógico es conceptual y abstracto. En esta etapa no se consideran aspectos sobre la implementación todavía, sólo se definen los tipos de información que se necesitan.

Una técnica comúnmente utilizada para modelar los requerimientos lógicos de información, es el modelado entidad-relación, en donde se identifican las cosas de importancia o de interés (entidades), las propiedades de estas cosas (atributos) y como se relacionan entre sí (relaciones)[10].

El diseño lógico busca arreglar los datos en series de relaciones lógicas llamadas entidades y atributos. Una entidad representa información de interés para el negocio y un atributo representa las características de una entidad. Cuando se transforma el diseño lógico en un diseño físico, las entidades se convierten en tablas y los atributos en columnas, y de la misma forma, las relaciones establecidas en las entidades se convierten en llaves foráneas.

Para asegurarnos que los datos son consistentes, además necesitamos identificadores únicos para cada registro u ocurrencia en cada entidad, lo cual comúnmente se conoce como llave primaria o llave única en el diseño físico.

A pesar de que el modelado entidad-relación tradicionalmente se ha asociado a modelos altamente normalizados para aplicaciones transaccionales, esta técnica sigue siendo válida para el diseño de almacenes de datos, para lo cual se realiza un modelado dimensional. En el modelado dimensional, en lugar de buscar unidades de información atómicas (entidades y atributos) y todas las relaciones entre ellas, debe identificarse cual información pertenece a un hecho central de análisis y cual información pertenece a una dimensión para ampliar dicho análisis.

[10] Data Modeling and Database Design , William Cobs, 1992.

El modelo lógico del almacén de datos debe dar como resultado un conjunto de entidades y atributos que se convertirán en tablas de hechos y dimensiones. Además debe ser un modelo de datos operacionales provenientes de varias fuentes orientado a algún tema específico de análisis del negocio.

1.5.1.1 Hechos y dimensiones en el modelo de estrella

El modelo lógico frecuentemente usado en un almacén de datos es el *modelo de estrella* representado por una entidad que concentra la información histórica de los *hechos* o acontecimientos que son el sujeto de análisis, y un conjunto de entidades circundantes llamadas *dimensiones* que son los aspectos bajo los cuales se desea visualizar los hechos.

Un ejemplo de modelo de estrella clásico es una entidad de hechos que refleja las ventas realizadas durante los últimos 5 años y 4 dimensiones al rededor de ella, por ejemplo una entidad que representa los productos que vende la organización, una que representa las regiones donde se venden los productos, una que representa los clientes y una que representa las fechas o representaciones de tiempo significativas para la organización (Trimestres, cuatrimestres, año fiscal etc.), en donde lo que se desea es realizar análisis de las ventas por producto, región, cliente y período.

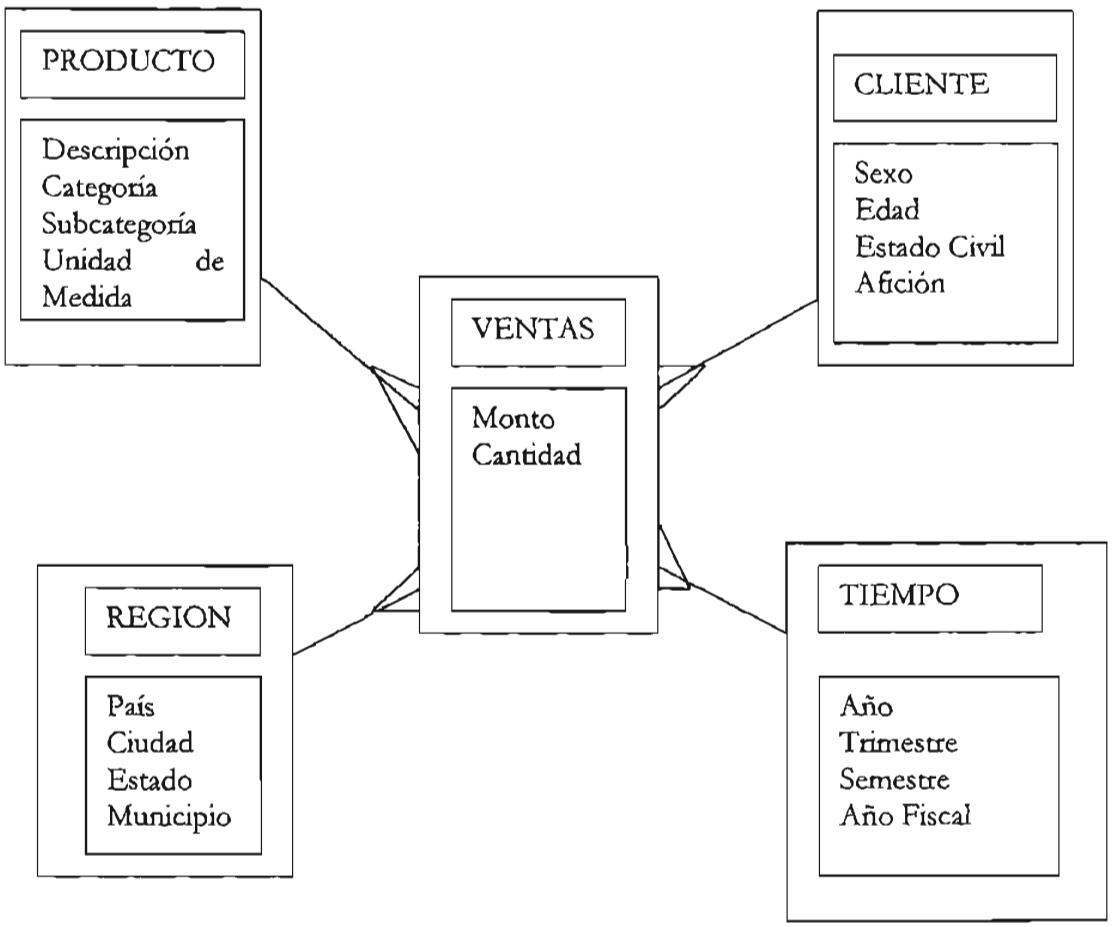


Figura 5 Ejemplo de modelo de estrella para un almacén de datos

En la figura 5 observamos que el tema central de análisis, son las ventas de una compañía, para lo cual se tendrá una tabla de hechos llamada VENTAS que concentrará los montos y cantidades de ventas realizadas a través del tiempo y el análisis estará orientado a visualizar la información bajo 4 diferentes dimensiones, PRODUCTOS vendidos, CLIENTES, REGIONES y unidades de TIEMPO (tales como trimestres, sexenios, años fiscales etc.), para lo cual se tendrán 4 tablas de dimensiones con sus atributos particulares.

1.5.1.2 Jerarquías

En este modelo también se pueden definir jerarquías, que son estructuras lógicas que determinan niveles de ordenamiento o de organización de los datos. Por ejemplo, en una dimensión de tiempo, una jerarquía podría agrupar datos de un nivel de detalle mensual a un nivel trimestral y ascendentemente a un nivel anual. Una jerarquía también se puede utilizar para definir una ruta de navegación y para establecer una estructura organizacional o familiar.

Dentro de una jerarquía, cada nivel está lógicamente conectado a niveles inferiores y superiores. Los valores de los datos a niveles menores se agrupan en valores de datos de niveles superiores. Una dimensión puede estar compuesta por más de una jerarquía.

Por ejemplo en la dimensión de "regiones", podría haber una jerarquía que establezca la manera en que la organización agrupa sus ventas territorialmente:

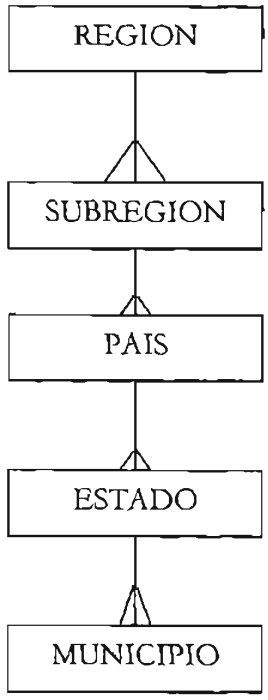


Figura 6 Ejemplo de un modelo jerárquico de regiones

Las jerarquías también agrupan niveles de lo general a lo granular. Las herramientas de consulta para almacenes de datos, usan las jerarquías para desplegar niveles de detalle de distinta granularidad.

Cuando se diseñan jerarquías, deben considerarse las relaciones entre las estructuras del negocio. Por ejemplo una organización divisional con multiniveles de ventas.

Las jerarquías imponen una estructura familiar a los valores de las dimensiones. Para un valor de nivel particular, un valor en el siguiente nivel superior es antecesor o padre y un valor en el siguiente nivel inferior es un sucesor o hijo dentro de la jerarquía.

Los manejadores de bases de datos con capacidades de soporte a almacenes de datos, permiten la definición de objetos tipo dimensiones y jerarquías, lo cual permite que los algoritmos internos del manejador identifiquen estos tipos de objetos y optimicen sus

planes de ejecución para el acceso a los datos, lo cual beneficia mucho los tiempos de respuesta del almacén de datos.

1.5.2 Diseño físico

Una vez finalizado el modelo lógico, se procede al diseño físico que establece la forma en que se organizará la información en la base de datos.

En esta etapa se realizan los mapeos de entidades a tablas, atributos a columnas, identificadores únicos a llaves primarias, relaciones a llaves foráneas etc.

También se analizan aspectos de volumen, crecimiento, depuración y reciclaje de los datos para determinar el tipo de estructuras físicas que se necesitarán implementar para soportar un modelo lógico y garantizar tiempos de respuesta aceptables.

Uno de los aspectos relevantes y críticos en el diseño físico de un almacén de datos es el volumen de información que se introducirá en el mismo, ya que esto es lo que determinará los requerimientos de hardware y software para soportar el sistema.

Cuando entramos en el análisis de volumen de información y nos encontramos con volúmenes considerables de datos a manejar (cientos de GIGABYTES o varios TERABYTES), debemos revisar las alternativas que tenemos para eficientar el uso de los recursos que tenemos y debemos recurrir a técnicas de almacenamiento y organización de la información que se encuentran implementadas intrínsecamente en el manejador y nos permitirán obtener los niveles de servicio y rendimiento deseados con costos de hardware minimizados.

Uno de los métodos intrínsecos en algunos manejadores de bases de datos que nos permitirán lograr lo expuesto en el párrafo anterior, son los métodos de particionamiento que se exponen y son el tema central de este trabajo.

Durante el diseño físico, al momento de convertir entidades en tablas (principalmente las de hechos, que normalmente son las que ocupan grandes volúmenes de datos), se debe determinar la forma más eficiente de particionar físicamente estos datos, de tal forma que lógicamente (y desde la perspectiva del usuario final) es un solo ente que representa el tema de interés de análisis del negocio, pero físicamente los datos se encuentran acomodados para que su acceso y explotación sean eficientes. Para esto debemos comprender el modelo lógico de datos, su razón de ser, la manera en que se desea explotar y el tipo de consultas que normalmente podrían realizarse, para así de acuerdo a los diversos métodos de particionamiento que existen, determinar el adecuado para un óptimo análisis de información y mantenimiento.

Capítulo 2

Métodos de particionamiento de datos

De acuerdo con lo expuesto en el capítulo 1, durante el diseño físico de un almacén de datos, se debe determinar la mejor manera de almacenar los datos considerando los siguientes aspectos:

- Conceptualmente la vista de los usuarios hacia los datos debe alinearse al modelo lógico y ser amigable; Es decir que la manera en que se almacenan físicamente los datos debe ser transparente para los usuarios finales.
- De acuerdo a consultas típicas a un almacén de datos, el acceso a los mismos debe ser eficiente.
- Los mecanismos de respaldo, depuración y reciclado de los datos deben ser eficientes.
- Debe haber un óptimo aprovechamiento de la arquitectura y del espacio disponible en los discos.
- El uso de los recursos de *hardware* debe minimizarse.

En este capítulo se presentan métodos de particionamiento de datos que permiten organizar físicamente los datos de tal forma que los aspectos mencionados anteriormente pueden cubrirse obteniendo beneficios en mantenimiento y eficiencia en los procesos de extracción, transformación, carga y explotación de la información del almacén de datos.

Los métodos de particionamiento de datos resuelven la problemática de soportar tablas muy grandes, permitiendo descomponer dichas

tablas en piezas pequeñas y manejables llamadas particiones, sin perder sus características definidas en el diseño lógico y conceptual. De tal forma que las instrucciones de manipulación de datos (SQL), no deben modificarse de ninguna manera cuando dichas instrucciones operan sobre tablas particionadas, lo cual brinda transparencia entre el modelo físico y el modelo lógico.

Por otra parte, al particionar ciertas tablas se obtiene el beneficio de que ciertas operaciones de mantenimiento, pueden hacerse sobre las particiones que son de interés en lugar de hacerse sobre la tabla completa, lo cual reduce considerablemente los tiempos de mantenimiento.

Cada partición debe tener exactamente los mismos atributos lógicos tales como nombres de las columnas, tipos de datos, reglas de integridad etc. , pero los atributos físicos tales como el espacio reservado, las extensiones máximas que puede tener, el archivo físico donde se almacenará etc., son atributos definidos de forma particular para cada partición, lo cual brinda la flexibilidad de mantenimiento y reciclado de la información contenida en la tabla.

La siguiente figura muestra la organización física de una tabla particionada en comparación con una tabla no particionada:

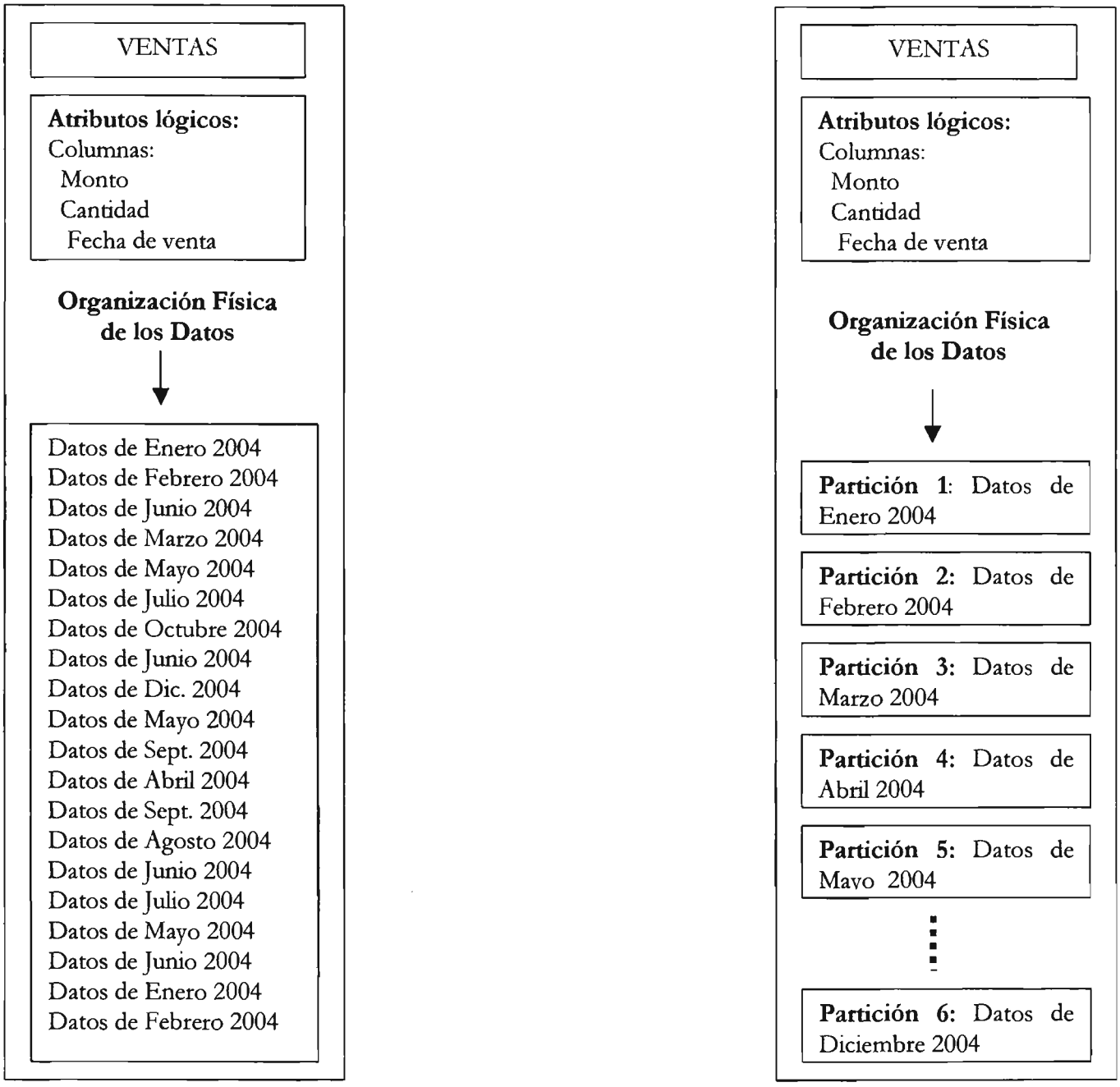


Figura 7 Comparación entre tabla particionada y no particionada

Partiendo del hecho de que algunos manejadores de bases de datos cuentan con capacidades para el manejo de particiones, se derivan los siguientes beneficios al aplicar métodos de particionamiento en dichos manejadores:

- Habilitan la realización de operaciones de transformación, carga y consultas a nivel de partición en lugar de realizarse a nivel de la tabla completa, derivando mejores tiempos de respuesta.

- Facilitan operaciones de respaldo, recuperación y reciclado ya que dichas operaciones pueden realizarse a nivel de partición.

- Los métodos de particionamiento pueden contemplarse desde la etapa de diseño de un sistema o bien puede implementarse en un sistema existente sin necesidad de hacer modificaciones a las unidades de programación.

- Reducen costos de recursos de cómputo y mantenimiento.

A continuación se explican tres métodos de particionamiento que podrían ser adecuados para la explotación y mantenimiento del almacén de datos del INEGI ya que como se analizará en el siguiente capítulo, el modelo de datos y las necesidades de explotación de la información, sugieren la utilización de estos métodos.

2.1 Método de particionamiento por rangos

Como su nombre lo indica, el método de particionamiento por rangos consiste en particionar los registros contenidos en una tabla de acuerdo a ciertos rangos definidos, los cuales normalmente son rangos de fechas, aunque también podrían ser rangos numéricos.

Para establecer un particionamiento por rangos sobre una tabla, se debe seleccionar una de sus columnas como la *llave de particionamiento* con la que se determina el criterio bajo el cual, los registros de la tabla se acomodarán físicamente en la partición que les corresponde.

Por ejemplo, en la tabla particionada que se muestra en la figura 7, se observa que las particiones agrupan datos de manera mensual, por lo que se puede seleccionar una columna de tipo fecha (FECHA_VENTA), y con ésta establecer el criterio de particionamiento, para que cada registro se acomode en la partición que le corresponde de acuerdo al valor que tenga en la columna FECHA_VENTA.

Como ya se ha mencionado, algunos manejadores de bases de datos relacionales, cuentan con capacidades intrínsecas para la definición de particiones extendiendo el lenguaje estándar de definición de datos ANSI SQL, de tal forma que dentro de la sintaxis de definición de una tabla se incluyen aspectos de particionamiento de la misma como se observa en el siguiente ejemplo:

```
CREATE TABLE VENTAS
(id_vendedor NUMBER(5),
nombre_vendedor VARCHAR2(30),
Monto_vendido NUMBER(10),
Fecha_venta DATE)
PARTITION BY RANGE(fecha_venta)
(
PARTITION ventas_enero2004 VALUES LESS THAN(TO_DATE('02/01/2004','DD/MM/YYYY')),
PARTITION ventas_febrero2002 VALUES LESS THAN(TO_DATE('03/01/2004','DD/MM/YYYY')),
PARTITION ventas_marzo2002 VALUES LESS THAN(TO_DATE('04/01/2004','DD/MM/YYYY')),
PARTITION ventas_abril2002 VALUES LESS THAN(TO_DATE('05/01/2004','DD/MM/YYYY'))
);
```

Como se puede observar en la instrucción SQL anterior, se define la tabla VENTAS con las columnas ID_VENDEDOR, NOMBRE_VENDEDOR, MONTO_VENDIDO y FECHA_VENTA con sus tipos de datos respectivos y se define un tipo de particionamiento por rangos estableciendo la columna FECHA_VENTA como la llave de particionamiento en donde los registros cuya FECHA_VENTA sea menor al 1 de febrero de 2002 serán ubicados en la partición VENTAS_ENERO2002, y los registros cuya FECHA_VENTA sea menor al 1 de marzo de 2002 serán ubicados en la partición VENTAS_FEBRERO2002, y así sucesivamente, de tal forma que al definir una partición por rangos se determina la columna que será la llave utilizada para particionar los datos, y los limites superiores de cada partición.

Como ya se mencionó anteriormente, el particionamiento por rangos no necesariamente tiene que realizarse por rangos de fechas, ya que dependiendo de las necesidades del sistema podría establecerse otro tipo de rangos numéricos sobre los cuales los registros se acomoden en las particiones que les correspondan. Por ejemplo, la tabla de ventas podría particionarse utilizando como *llave de particionamiento* la columna MONTO_VENDIDO y acomodar los datos de acuerdo a rangos de ventas como se muestra en el siguiente ejemplo:

```
CREATE TABLE VENTAS
(id_vendedor  NUMBER(5),
nombre_vendedor VARCHAR2(30),
Monto_vendido  NUMBER(10),
Fecha_venta    DATE)
PARTITION BY RANGE(monto_vendido)
(
PARTITION ventas_menores_100mil VALUES LESS THAN(100000),
PARTITION ventas_menores_1millon VALUES LESS THAN(TO_DATE(1000000),
PARTITION ventas_menores_10millones VALUES LESS THAN(10000000),
PARTITION ventas_menores_100millones VALUES LESS THAN(100000000)
);
```

En los ejemplos anteriores se observa que cada partición define su límite superior, y su límite inferior queda definido por la partición que le antecede.

2.2 Método de particionamiento por lista

En el método de particionamiento por lista, como su nombre lo indica, se determina una lista de valores válidos que la llave de particionamiento (una columna de la tabla) debe tener para cada partición.

Este método es utilizado cuando se desea organizar conjuntos de datos que no cumplen con un orden o no están relacionados de una forma específica.

Por ejemplo cuando la tabla de VENTAS se desea particionar por regiones geográficas, se puede tomar como llave de particionamiento una columna que determine la región donde fue realizada la venta y agrupar los datos en particiones que contengan datos de acuerdo a los puntos cardinales como se muestra en la siguiente definición:

```
CREATE TABLE VENTAS
(id_vendedor  NUMBER(5),
nombre_vendedor VARCHAR2(30),
Monto_vendido  NUMBER(10),
Fecha_venta   DATE,
Lugar_venta VARCHAR2(30))
PARTITION BY LIST (lugar_venta)
(
PARTITION ventas_norte VALUES ('NUEVO LEON','CHIHUAHUA','COAHUILA'),
PARTITION ventas_CENTRO VALUES ("DF,'JALISCO','PUEBLA'),
PARTITION ventas_SUR VALUES ('TABASCO','CHIAPAS','YUCATAN'),
);
```

Como se observa en el ejemplo anterior se define una tabla particionada por el método de lista, tomando como llave de particionamiento la columna LUGAR_VENTA y definiendo tres particiones, una para las ventas realizadas al norte del país, otra para las realizadas en el centro del país y otras realizadas al sur del país, suponiendo que el negocio realiza sus ventas en los estados mencionados para cada partición y en caso de abrir una sucursal nueva, sea por ejemplo "BAJA CALIFORNIA", se requeriría una operación de mantenimiento solamente en la partición VENTAS_NORTE para incluir en su lista de valores el estado mencionado.

2.3 Método de particionamiento por función hash

Este método es utilizado cuando no se sabe de antemano como será la distribución de los datos entre las particiones o no se tiene un criterio que distribuya los datos equilibradamente con posibilidades de tener unas particiones con grandes volúmenes de datos y otras con volúmenes muy pequeños lo cual dificulta la administración y mantenimiento de las mismas.

Es decir que cuando es difícil establecer una llave que determine un balanceo natural de los datos contenidos en cada partición, se puede utilizar el método de particionamiento por función hash que consiste en aplicar una función de hash a una columna numérica seleccionada como la llave de la partición, de tal forma que al aplicar la función de hash a la columna seleccionada, el valor obtenido es el que determina la partición a la que pertenecerá cada registro. La función hash se implementa automáticamente por el manejador de base de datos y no requiere una programación especial.

De esta manera sólo es necesario establecer cuántas particiones se van a manejar para la tabla y la función de hash aplicada a la llave de particionamiento determinará a cual de dichas particiones pertenecerá cada registro.

El siguiente ejemplo muestra como se define una tabla con particionamiento por función hash:

```
CREATE TABLE VENTAS
(id_vendedor  NUMBER(5),
Nombre_vendedor VARCHAR2(30),
Monto_vendido  NUMBER(10),
Numero_semana       NUMBER(2))
PARTITION BY HASH(id_vendedor)
PARTITIONS 4 ;
```

En el ejemplo anterior se define la tabla VENTAS con particionamiento por hash, tomando la columna ID_VENDEDOR como la llave de particionamiento a la cual se le aplicará la función de hash y dependiendo del valor resultante, el registro se ubicará en una de las 4 particiones definidas en la instrucción.

2.4 Subparticionamiento y paralelismo

Cuando se aplica un tipo de particionamiento y se obtiene como resultado un conjunto de particiones que en sí mismas son muy grandes, se tiene la opción de *subparticionar* cada una de las particiones resultantes nuevamente, teniendo como resultado *subparticiones* manejables que permitirán mayor eficiencia en algunas operaciones y además habilitarán la opción de paralelizar operaciones a nivel partición obteniendo tiempos de respuesta reducidos.

Por ejemplo se puede tener el caso de la tabla de VENTAS particionada por el método de lista agrupando los datos por regiones y tener como resultado particiones regionales independientes con volúmenes de información grandes, mismas que a su vez pueden ser particionadas por el método de hash, digamos con 4 subparticiones por partición regional las cuales serán de tamaños uniformes, de tal forma que una operación realizada sobre la partición de una región puede paralelizarse con un grado de paralelismo de 4, dado que la partición esta subparticionada en 4 y así la operación se realiza en menor tiempo dado que al paralelizar se genera un proceso por cada subpartición para realizar la tarea deseada.

La siguiente instrucción muestra la definición de una tabla particionada por el método de lista y a su vez subparticionada por el método hash:

```
CREATE TABLE VENTAS
```

(id_vendedor NUMBER(5),
nombre_vendedor VARCHAR2(30),
Monto_vendido NUMBER(10),
Fecha_venta DATE,
Lugar_venta VARCHAR2(30))
PARTITION BY LIST (lugar_venta)
SUBPARTITION BY HASH (id_vendedor)
Partitions 4
(
PARTITION ventas_norte VALUES ('NUEVO LEON','CHIHUAHUA','COAHUILA'),
PARTITION ventas_CENTRO VALUES ('DF','JALISCO','PUEBLA'),
PARTITION ventas_SUR VALUES ('TABASCO','CHIAPAS','YUCATAN'),
);

Otro ejemplo de subparticionamiento o particionamiento compuesto sería el siguiente:

```
CREATE TABLE VENTAS
(id_vendedor  NUMBER(5),
 monto_vendido  NUMBER(10),
 Fecha_venta    DATE)
PARTITION BY RANGE(fecha_venta)
SUBPARTITION BY HASH(id_vendedor)
Partitions 4
(PARTITION              ventas_ene2000          VALUES          LESS
THAN(TO_DATE('02/01/2000','DD/MM/YYYY'))
 PARTITION              ventas_feb2000          VALUES          LESS
THAN(TO_DATE('03/01/2000','DD/MM/YYYY'))
 PARTITION              ventas_mar2000          VALUES          LESS
THAN(TO_DATE('04/01/2000','DD/MM/YYYY'))
 PARTITION              ventas_abr2000          VALUES          LESS
THAN(TO_DATE('05/01/2000','DD/MM/YYYY'))
 PARTITION              ventas_may2000          VALUES          LESS
THAN(TO_DATE('06/01/2000','DD/MM/YYYY')));
```

En el ejemplo anterior se crea la tabla VENTAS particionada por rangos mensuales usando como llave de particionamiento la columna FECHA_VENTA y cada partición mensual se subparticiona por hash generándose 4 subparticiones por cada partición mensual. Como se explicó anteriormente, dependiendo de la naturaleza de los datos del negocio, es como se toma la decisión de establecer el método de particionamiento y subparticionamiento. En el ejemplo anterior se puede estar en el caso en que las ventas mensuales son muchas y aún a nivel mensual se obtienen particiones muy grandes por lo que se decide subparticionar cada partición mensual por el método de hash.

Capítulo 3

Caso de Estudio

En este capítulo se presenta la síntesis metodológica del censo general de población y vivienda del Instituto Nacional de Estadística, Geografía e Informática (INEGI), realizado en el año 2000.

En esta síntesis se exponen las características del censo, de las cuales, se derivan requerimientos de información y modelos de datos, los cuales serán analizados en el siguiente capítulo para determinar los métodos de particionamiento adecuados para las grandes tablas que conforman el modelo de datos del almacén que contiene la información recolectada en el censo.

3.1 El XII Censo General de Población y Vivienda 2000

El XII Censo de población y vivienda 2000 es un proyecto de generación de estadística que realizó el INEGI, en el cual se captó información sobre las características sociodemográficas de la población en México.

A continuación se describen los aspectos generales del censo.

3.1.1 Objetivo

El censo de población y vivienda tiene como objetivo general proporcionar información estadística indispensable para el análisis y

la evaluación de la composición, distribución y el crecimiento de la población y las viviendas en México.

Para el XII censo general de población y vivienda 2000 se establecieron los siguientes objetivos particulares:

- Generar información demográfica y socioeconómica sobre el país.

- Asegurar la máxima desagregación geográfica de la información. Es decir, poder utilizar y analizar la información a los máximos niveles geográficos de detalle.

- Enriquecer las series históricas de datos estadísticos, manteniendo en lo posible la comparabilidad nacional e internacional.

- Construir marcos de muestreo para encuestas.

Para dar cumplimiento a estos objetivos, el Censo 2000 contó con un marco legal sustentado en la Constitución Política de los Estados Unidos Mexicanos, la ley de Información y Estadística y Geográfica y el decreto presidencial emitido el 1 de diciembre de 1999 y publicado en el Diario Oficial de la Federación, en el cual es entonces presidente Constitucional de los Estados Unidos Mexicanos, Dr. Ernesto Cedillo Ponce de León, declaró de interés nacional el proceso de preparación, organización, levantamiento y publicación del XII Censo General de población y vivienda 2000.

La base jurídica del Censo 2000 garantiza la confidencialidad de la información censal, es decir, obliga al INEGI a guardar en estricta reserva los datos individuales y a utilizarlos en forma agrupada,

únicamente con fines estadísticos; además, enfatiza el deber de la población de proporcionar los datos solicitados por los funcionarios censales.

3.1.2 Antecedentes

El XII Censo General de Población y Vivienda 2000 da continuidad a una larga tradición censal mexicana que se remonta a 1895, año en que se realizó el primer censo de población de la época moderna en México

Desde entonces se han levantado 12 censos de población, mismos que se han realizado cada década en los años terminados en cero, con excepción de 1920, cuando por razones políticas y sociales tuvo que posponerse hasta 1921. Además, entre 1990 y 2000 se realizó un recuento intercensal denominado Conteo de Población y Vivienda 1995.

A partir de 1950 se realizan los censos de población y de vivienda en forma simultánea, ya que en los anteriores sólo se captaban algunas características de la vivienda.

A continuación se describen las principales características de los censos de población y vivienda en México de 1895 a 2000:

Los censos de población de 1895 y 1900 fueron de hecho y derecho, con tres tipos de cuestionarios, por vivienda y autoempadronamiento.

De hecho o de facto, es aquel en donde se enumera a la población en el lugar en que se encuentre presente en el momento del levantamiento

En el de derecho o de jure se enumera a la población en el lugar en que vive normalmente, es decir, en su lugar de residencia habitual.

Los censos de 1910 y 1921 también fueron censos de hecho, con un cuestionario por vivienda y autoempdronamiento.

Los censos de 1930 a 1960 fueron de derecho con un cuestionario colectivo y se realizaron a través de la entrevista directa.

En 1970, 1980 y 1990, los censos fueron de derecho; se aplicó un cuestionario por vivienda mediante entrevista directa

El de 2000 también fue un censo de derecho, se utilizaron dos tipos de cuestionarios, se aplico uno por vivienda por medio de la entrevista directa

3.1.3 Importancia

Por el nivel de desagregación geográfica con que son presentados sus resultados, el censo es un recurso indispensable para realizar estudios acerca de la situación actual del país, para reconocer los avances y rezagos en los niveles de bienestar de la población, así como para apoyar las diversas tareas que llevan a cabo los distintos sectores de la sociedad.

3.1.4 Unidades de análisis

Las dos principales unidades de análisis del XII Censo General de población y Vivienda 2000 fueron los residentes habituales y las viviendas.

Se consideró como residente habitual a toda persona que habita normalmente en la vivienda, esto es, que en ella duerme, prepara sus alimentos, come y se protege del ambiente, y por ello la reconoce como su lugar de residencia.

Como vivienda se consideró a todo espacio delimitado normalmente por paredes y techos, de cualquier material, con entrada independiente; que se utiliza para vivir, esto es, dormir, preparar los alimentos, comer y protegerse del ambiente.

3.1.5 Variables

Las variables que captó el Censo caracterizan a la población residente en México respecto a su composición sociodemográfica, las cuales responden a necesidades especificas de información en el campo de estudio.

Temática censal

Con base en las consultas, el análisis, las discusiones y las pruebas realizadas se determinaron los temas para el XII Censo General de Población y Vivienda 2000, los que se agruparon en tres grandes bloques: vivienda, número de residentes y de hogares, y características demográficas, sociales, educativas y económicas de la población

A continuación se presentan las variables captadas por categorías generales de estudio y tema:

3.1.5.1 Viviendas

Tipo y clase de vivienda.

Materiales de construcción en paredes, techos y recubrimiento del piso.

Disponibilidad de espacios, total de cuartos, cuartos dormitorios y cocina.

Disponibilidad de frecuencia del servicio de agua entubada.

Disponibilidad y exclusividad de servicio sanitario y conexión de agua.

Disponibilidad de drenaje y electricidad.

Combustible utilizado para cocinar.

Tenencia de la vivienda.

Antigüedad de la vivienda.

Eliminación de basura.

Bienes en la vivienda.

Numero de hogares.

3.1.5.2 Hogares

Total de residentes

Gasto común

3.1.5.3 Población

Sexo, edad, y relación de parentesco de los integrantes del hogar con él (la) jefe(a) del mismo.

Fecundidad y mortalidad: número de hijos fallecidos, hijos sobrevivientes, fecha de nacimiento del último hijo nacido vivo y, de éste, sobrevivencia; y en su caso, edad al morir.

Migración : lugar de nacimiento, lugar de residencia en 1995 (estado o país y municipio o delegación) y causa de la migración.

Migración internacional: el censo captó la migración de las personas que se fueron a vivir a otro país entre enero de 1995 y el momento del levantamiento, y distingue a los migrantes que aún viven en otro país y a los que ya regresaron. De esta manera, se ofrece información sobre, sexo, edad, lugar de origen, fecha de nacimiento, país de destino, país de residencia y fecha de retorno.

3.1.5.4 Características sociales

- Étnicas: población hablante de lengua indígena, condición de habla española, tipo de lengua y pertenencia étnica.
- Religión.
- Servicio de salud: derechohabiencia y uso de servicios de salud.
- Discapacidad: tipo y causa de la discapacidad.
- Estado conyugal.

3.1.5.5 Características educativas

- Alfabetismo: asistencia escolar, causa de abandono escolar, nivel de instrucción, antecedente escolar y nombre de la carrera.

3.1.5.6 Características económicas

- Condición de actividad, ocupación principal, situación en el trabajo, sector de actividad, ingreso por trabajo, horas trabajadas, prestaciones laborales, lugar de trabajo(municipio o delegación, entidad o país) y otros ingresos.

Cabe hacer mención que respecto al XI censo General y Vivienda 1990, las variables adicionales aplicadas a toda la población y vivienda que se incluyen en el censo 2000, son:

- Bienes en la vivienda
- Uso exclusivo del sanitario
- Derechohabiencia o servicio de salud
- Tipo de discapacidad
- Municipio de residencia en 1995
- Antecedente escolar
- Nombre de la carrera
- Hijos fallecidos
- Fecha de nacimiento y condición de supervivencia, en su caso edad al morir
- Verificación de actividad económica

Las variables adicionales incluidas en el cuestionario ampliado, son:

- Dotación de agua
- Antigüedad en la vivienda
- Causa de discapacidad
- Uso de servicio de salud
- Causa de emigración
- Causa de abandono escolar
- Pertenencia étnica
- Prestaciones laborales
- Lugar de trabajo
- Otros ingresos

Así como todo un apartado de migración internacional, cuyas variables son:

- Condición de migración internacional.
- Número de personas
- Personas migrantes
- Lista de personas
- Condición de residencia
- Sexo
- Edad
- Lugar de origen
- Fecha de emigración
- País de destino
- País de residencia
- Fecha de retorno

3.1.6 Principales productos

Como productos principales se tienen los Tabulados Básicos del XII Censo General de Población y Vivienda 2000, a nivel nacional para cada entidad federativa, con desglose municipal, además de la Síntesis de Resultados del XII censo General de Población y Vivienda 2000 de algunas entidades, así como los resultados de la muestra censal, en una publicación con tabulados de indicadores derivados del cuestionario amplio, con desglose estatal y por tamaño de localidad.

3.1.6.1 Difusión de resultados

La estrategia para la difusión de resultados consideró tres etapas. En la primera se publicaron los resultados preliminares en un producto impreso, en el que se integró información de la población total, su distribución por sexo, la densidad de población, el número de viviendas particulares y de ocupantes por vivienda. Estos datos se difundieron a ocho meses del levantamiento y son producto de la muestra censal. Consiste en una publicación de tabulados que muestran indicadores derivados del cuestionario amplio, con desglose estatal y por tamaño de localidad.

En la segunda etapa, después de un año de concluido el levantamiento del censo, se publicaron los resultados definitivos, concebidos como la producción básica del censo (tabulados básicos).

También se cuenta con la publicación de integración territorial (Iter) por entidad federativa, donde aparecen los principales resultados para cada una de las localidades de la entidad.

Por otra parte, se elaboró una publicación que contiene gráficas y mapas con estatificaciones de una selección de los principales indicadores censales.

Esta segunda etapa se complementa con sistemas para la consulta de información en disco compacto como el Sistema de Consulta de Información Censal(Scince) y el Sistema de Consulta de Tabulados, los cuales permiten al usuario, en el primer caso, relacionar información estadística con el espacio geográfico al que pertenece, y en el segundo, se puede tener acceso a los tabulados básicos y manipularlos. Además, están disponibles en disco compacto los archivos de la muestra del censo, con el fin de que el usuario pueda generar los indicadores que resulten de su interés.

Finalmente, en una tercera etapa que inició en el año 2001, se elaboraron publicaciones que presentan una descripción de la temática censal desde un punto de vista sociodemográfico, incluyendo un comparativo con información de censos anteriores. Estos productos están estructurados en tres niveles: el geográfico (nacional, estatal, zonas metropolitanas, entre otros), y el de subpoblaciones (jóvenes discapacitados etc.).

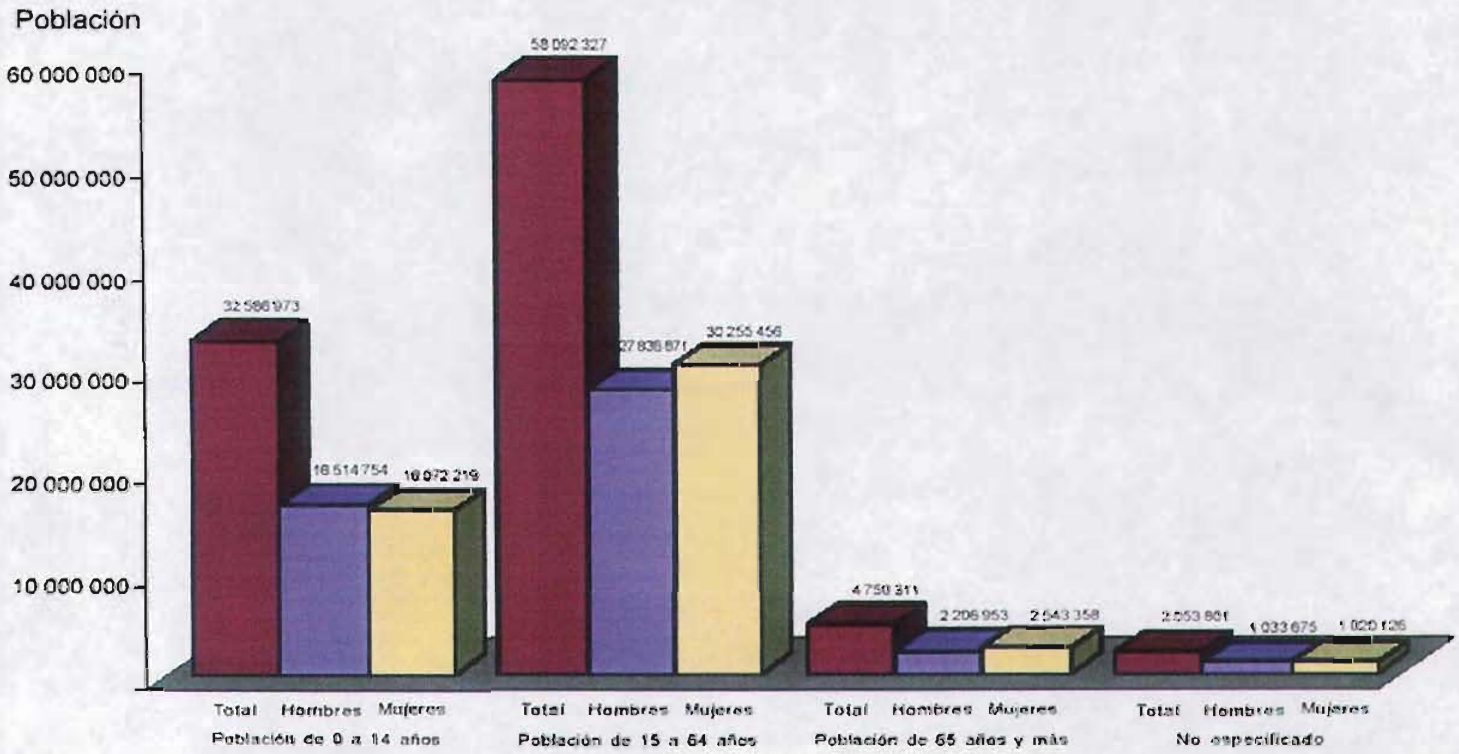
Además, en esta tercera etapa se planteó incluir una serie de publicaciones temáticas con tabulados comparativos, con el fin de ampliar la disponibilidad de la información captada por el XII Censo General de Población y Vivienda 2000. Dichos ejemplares pueden adquirirse o consultarse en el área de difusión del mismo instituto.

3.1.8 Ejemplo de resultados

En los siguientes cuadros y gráficas se presentan algunos resultados del XII Censo General de Población y Vivienda 2000, según las variables de población, fecundidad, mortalidad, migración, lengua indígena, religión, educación, servicio de salud, discapacidad, estado conyugal, empleo, hogares y viviendas.

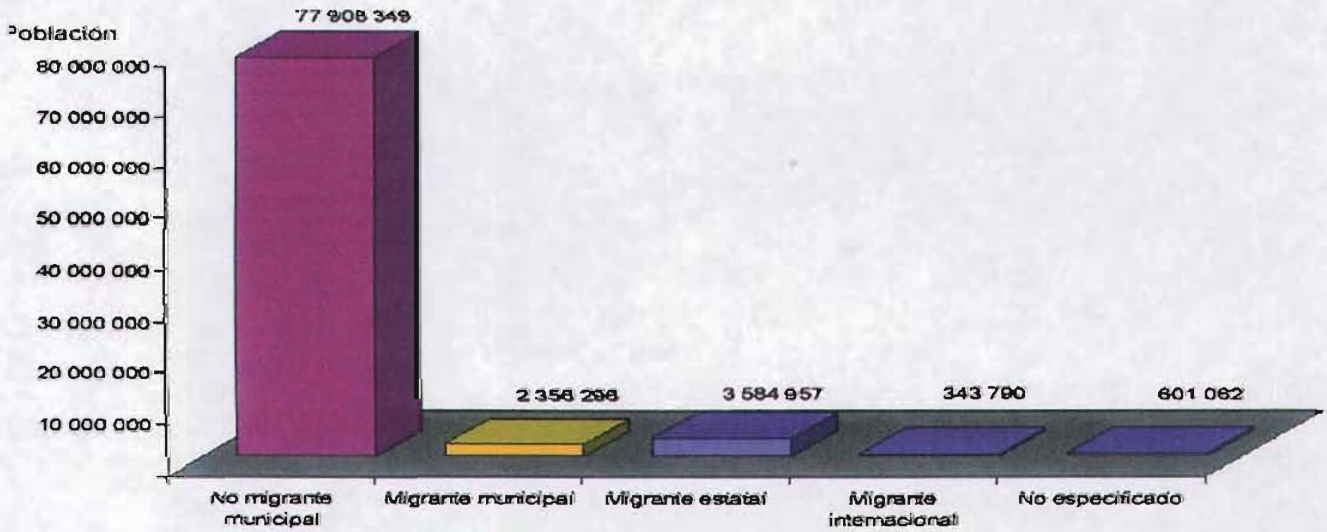
La fuente de donde se tomaron los datos para las gráficas fueron los Tabulados Básico Estados Unidos Mexicanos, tomo I, II y III del XII censo General de Población y Vivienda 2000, publicado por el Instituto Nacional de Estadística, Geográfica e Informática.

Estados Unidos Mexicanos
Población y su distribución, según grandes grupos de edad y sexo



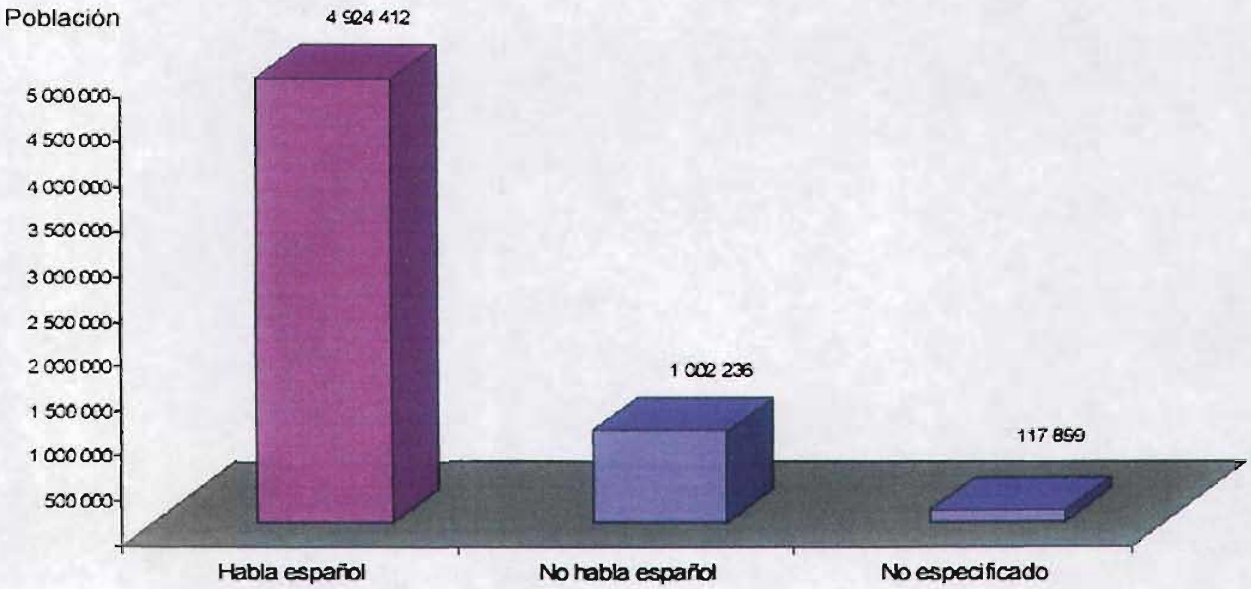
Gráfica 1 Población y su Distribución

Estados Unidos Mexicanos
Distribución de la población de 5 años y más, según condición migratoria municipal, estatal e internacional



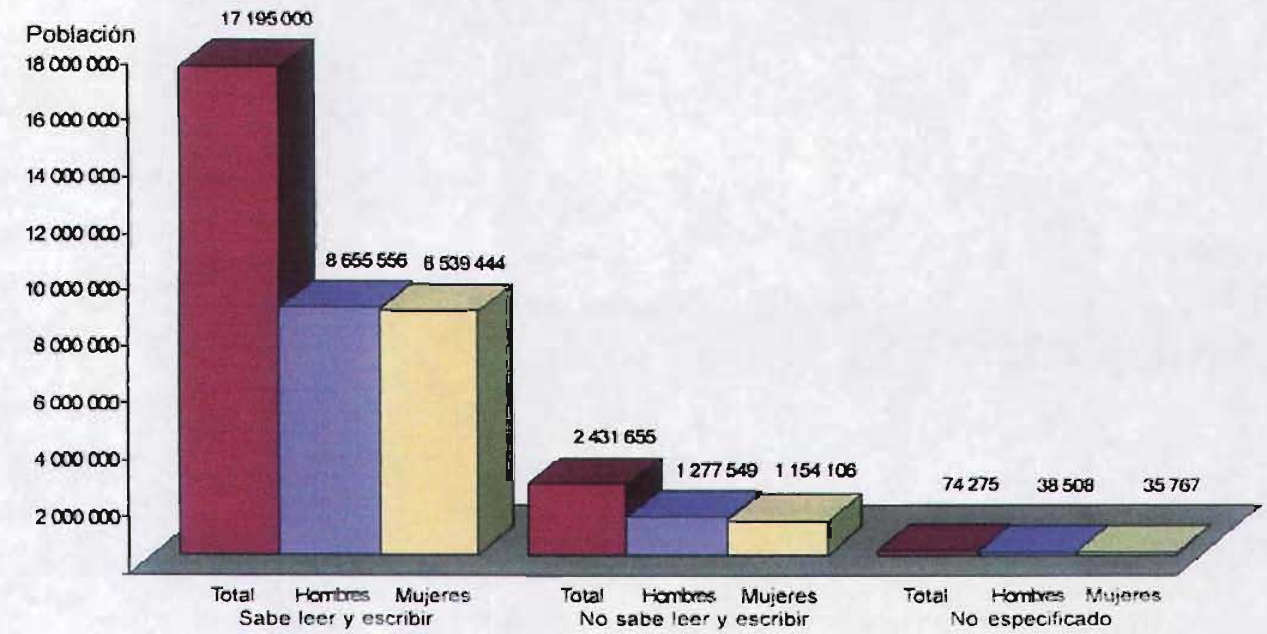
Gráfica 2 población de 5 años y más

Estados Unidos Mexicanos
Población de 5 años y más que habla lengua indígena, según condición de habla indígena y habla española



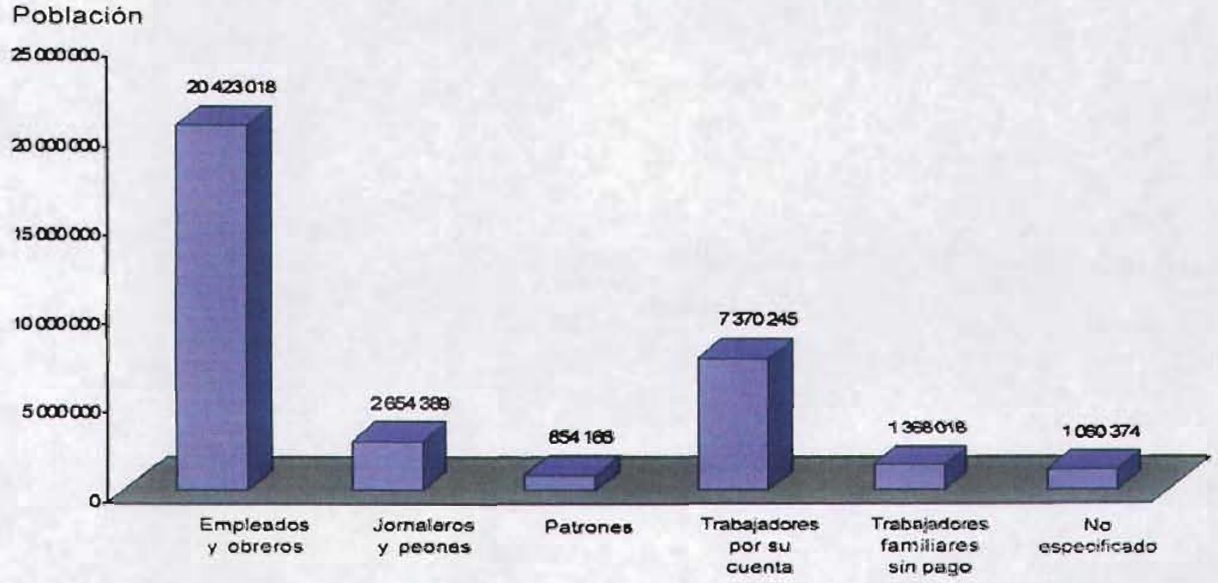
Gráfica 3 Población que habla lengua indígena y habla española

Estados Unidos Mexicanos
Población de 6 a 14 años por sexo, según aptitud para leer y escribir



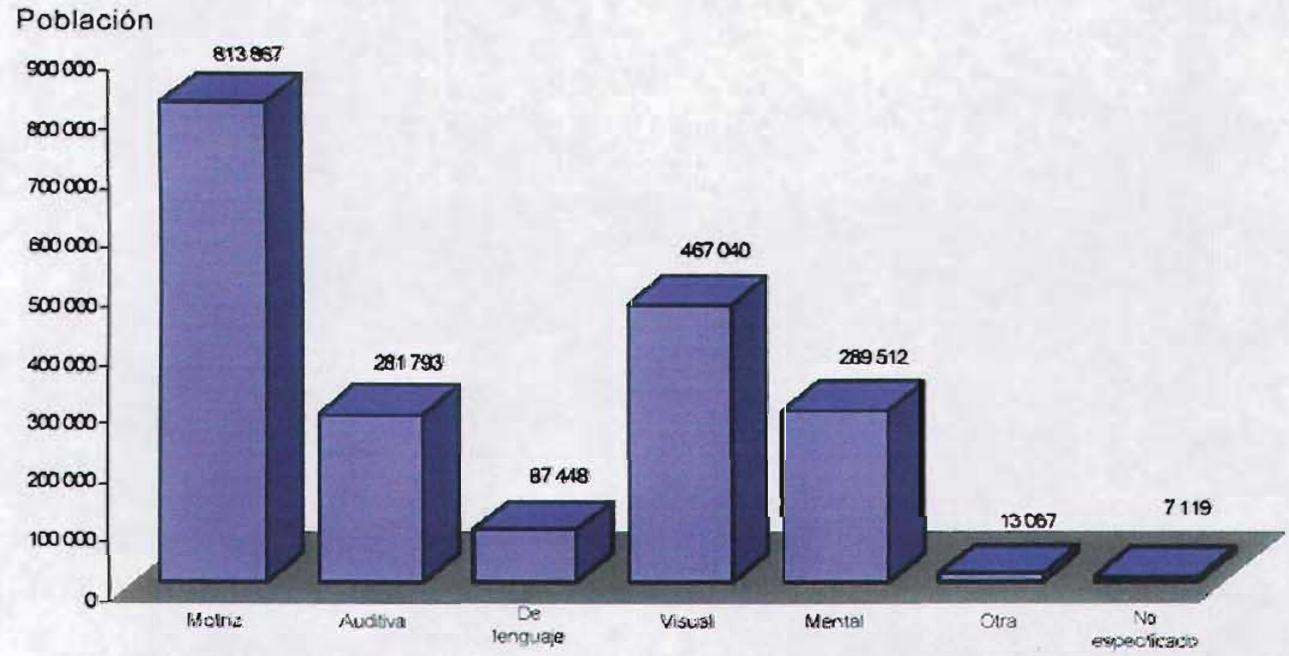
Gráfica 4 Población de 6 a 14 años por sexo

Estados Unidos Mexicanos
Distribución de la población ocupada, según situación en el trabajo



Gráfica 5 Distribución de población ocupada

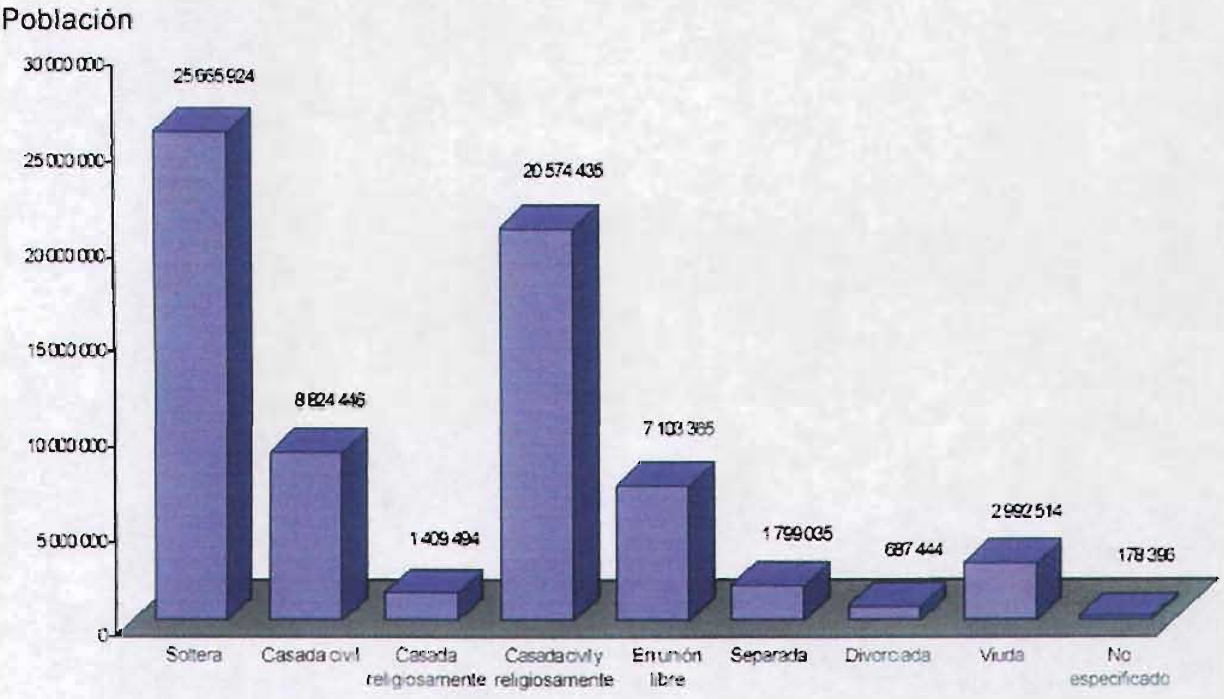
Estados Unidos Mexicanos
Población discapacitada, según tipo de discapacidad



NOTA: La suma de los distintos tipos de discapacidad es mayor al total, por aquella población que presenta más de una discapacidad.

Gráfica 6 Población discapacitada

Estados Unidos Mexicanos
Población de 12 años y más, según su estado conyugal



Gráfica 7 Población de 12 años y más y estado conyugal

3.2 Modelos de datos

Después del análisis de las características del censo, se determinaron los requerimientos de información y se desarrolló el modelos de datos que se presenta en el ANEXO A, el cuál se explica a continuación [11].

El modelo de datos tiene tres tablas que representan los temas centrales del almacén de datos que se podrían considerar como los hechos en el modelo de estrella explicado en capítulos anteriores. Dichas tablas son VIVIENDAS, HOGARES y POBLADORES, las cuales están relacionadas entre sí mediante llaves foráneas representadas por líneas que conectan estas tablas en una relación "uno a muchos" que se describe a continuación:

Una vivienda contiene uno o mas hogares.
Un hogar contiene uno o mas pobladores.

La tabla de VIVIENDAS representa el centro de una estrella rodeada de pequeños catálogos que representan las dimensiones bajo las cuales puede obtenerse conocimiento acerca de las viviendas censadas.

En la siguiente tabla se describen las dimensiones relacionadas con las viviendas:

DIMENSION (Catálogo)	**Descripción**
TAMANIOS_LOC	Contiene los tipos de tamaño de localidad a los que una vivienda

[11] Para mayor información consultar www.inegi.gob.mx

	puede pertenecer.
BIENES_VIVIENDA_1	Contiene los tipos de RADIO, AUTO y REFRIGERADOR que una vivienda puede tener.
BIENES_VIVIENDA_2	Contiene los tipos de LAVADORA, COMPUTADORA y TELEVISOR que una vivienda puede tener.
BIENES_VIVIENDA_3	Contiene los tipos de VIDEO y TELEFONO que una vivienda puede tener.
BIENES_VIVIENDA_4	Contiene los tipos de LICUADORA y BOILER que una vivienda puede tener.
GASTOS_COMUNES	Contiene los tipos de gastos comunes que una vivienda puede tenener.
CLASES_VIVIENDA	Contiene las clases de vivienda que se pueden tener.
MATERIAL_PAREDES	Contiene los tipos de materiales de las paredes que una vivienda puede tener.
MATERIAL_TECHOS	Contiene los tipos de material de los techos que una vivienda puede tener.
MATERIAL_PISOS	Contiene los tipos de material de piso que una vivienda puede tener.

COCINAS	Contiene los tipos de cocina que una vivienda puede tener.
CUARTOS_DORMITORIOS	Contiene rangos de cantidades de dormitorios que una vivienda puede tener.
DOTACIÓN_AGUA_DIAS	Contiene rangos de días de dotación de agua que una vivienda puede tener.
SERVICIOS_SANITARIOS	Contiene los tipos de servicios sanitarios que una vivienda puede tener.
CONEXIONES_AGUA	Contiene los tipos de conexión de agua que una vivienda puede tener.
SERVICIO_DRENAJE	Contiene los tipos de servicio de drenaje que una vivienda puede tener.
SERVICIO_ELECTRICIDAD	Contiene los tipos de servicio de electricidad que una vivienda puede tener.
COMBUSTIBLES	Contiene los tipos de combustible que una vivienda puede usar.
TENENCIAS_PROP	Contiene los tipos de tenencia de propiedad que una vivienda puede tener.
ELIMINACIONES_BASURA	Coniene los tipos de eliminación de basura que una vivienda puede tener.
RECOLECCIONES_BASURA	Contiene los tipos de recolección de basura que una vivienda puede tener.

La tabla de HOGARES representa el centro de otra estrella rodeada de pequeños catálogos que representan las dimensiones bajo las cuales puede obtenerse conocimiento acerca de los hogares censados.

En la siguiente tabla se describen las dimensiones relacionadas con los hogares:

DIMENSION (Catálogo)	**Descripción**
CONDICION_MIGRANTES	Contiene los tipos de condición de migrantes que un hogar puede tener.
TIPOS_HOGARES	Contiene los tipos de hogares que se pueden tener.
LOCALIDADES	Contiene los tipos de localidades a las que un hogar puede pertenecer.

La tabla de POBLADORES representa el centro de otra estrella rodeada de pequeños catálogos que representan las dimensiones bajo las cuales puede obtenerse conocimiento acerca de los pobladores censados.

En la siguiente tabla se describen las dimensiones relacionadas con los pobladores:

DIMENSION (Catálogo)	**Descripción**
CAUSAS_EMIGRACIÓN	Contiene los tipos de causas de emigración que un poblador puede tener.
OCUPACIONES	Contiene los tipos de ocupación que un poblador puede tener.
ACTIVIDADES	Contiene los tipos de

	actividades que un poblador puede tener.
SITUACIONES_TRABAJO	Contiene los tipos de situación de trabajo que un poblador puede tener.
VACACIONES_PAGADAS	Contiene los tipos de vacaciones pagadas que un poblador puede tener.
REPARTO_UTILIDADES	Contiene los tipos de reparto de utilidades que un poblador puede tener.
SAR_AFORE	Contiene los tipos de SAR o AFORE que un poblador puede tener.
SERVICIOS_MEDICOS	Contiene los tipos de servicio medico que un poblador puede tener.
CARRERAS	Contiene los tipos de carrera que un poblador puede tener.
ESCOLARIDADES	Contiene los tipos de escolaridad que un poblador puede tener.
NIVELES_ACADÉMICOS	Contiene los tipos de nivel académico que un poblador puede tener.
ANTECEDENTES_ESCOLARES	Contiene los tipos de antecedentes escolares que un poblador puede tener.
ESTADOS_CONYUGALES	Contiene los tipos de estado conyugal que un poblador puede tener.
PARENTESCOS	Contiene los tipos de

	parentesco que un poblador puede tener.

Capítulo 4

Aplicación Práctica

Con la información presentada en los capítulos anteriores, se procede al análisis del caso de estudio para aplicar los conceptos y métodos ya expuestos.

Primordialmente se busca determinar los métodos de particionamiento adecuados para las tablas que conforman el modelo físico del almacén de datos.

Bajo el contexto que brinda el caso de estudio y observando el modelo de datos, se observa que los temas centrales del almacén de datos son:

a) Las viviendas
b) Los hogares
c) Los pobladores

Estos temas centrales se ven representados en el modelo de datos por las tablas TR_VIVIENDAS, TR_HOGARES y TR_POBLADORES las cuáles se identifican como los *hechos* dentro del contexto del almacén de datos.

Se puede asumir que dichas tablas son las que tienen el mayor volumen de información, por lo tanto serán el punto central de este análisis para determinar el método adecuado de particionamiento.

Dado que cualquier método de particionamiento depende de las columnas que conforman la tabla a particionar, se observan las columnas de cada tabla y bajo el contexto que brinda el caso de estudio se determinan las columnas que servirán como llave de particionamiento.

Análisis de viviendas

Le estructura de la tabla TR_VIVIENDAS se presenta a continuación:

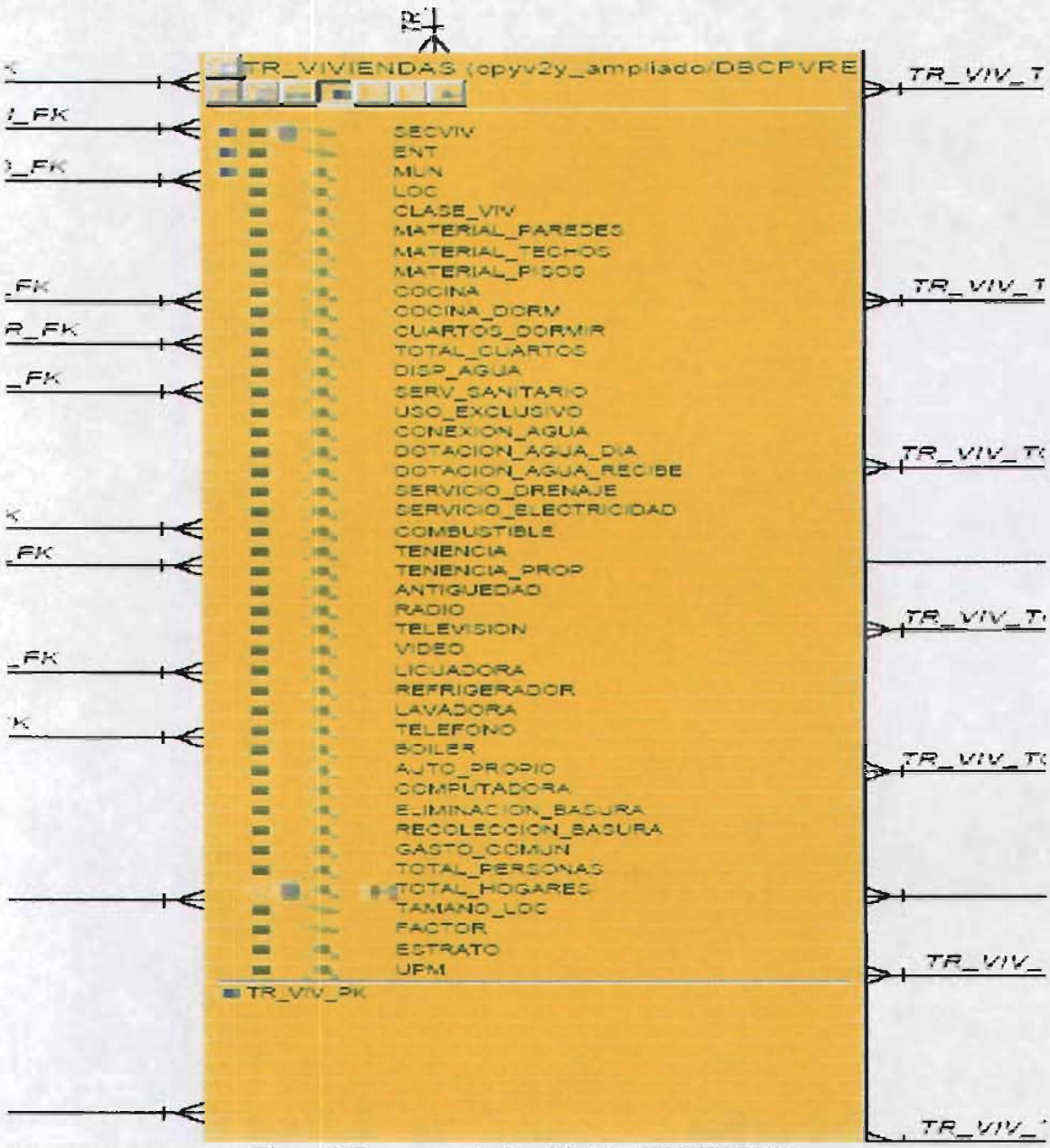


Figura 8 Estructura de la tabla TR_VIVIENDAS

Las líneas que se conectan con ésta tabla son llaves foráneas que representan las relaciones o dimensiones que existen alrededor de las viviendas sobre las cuáles se analiza la información recolectada en el censo, tales como catálogos de clases de vivienda, tipo de tenencia, tipo de toma de agua, tipo de servicio de drenaje etc.

De la estructura anterior y tomando como referencia la manera en que se desea analizar la información como se establece en el caso de estudio, las columnas que son útiles como llaves de particionamiento son:

COLUMNA	**DESCRIPCIÓN**
SECVIV	Número identificador de la vivienda
ENT	Entidad Federativa donde se ubica la vivienda.
MUN	Municipio donde se ubica la vivienda.
LOC	Localidad donde se ubica la vivienda.
CLASE_VIV	Clase de vivienda

Tabla 1 Columnas que pueden ser llave de particionamiento

Para determinar cual de las columnas mencionadas es adecuada para particionar la tabla TR_VIVIENDAS, se considera del caso de estudio que "*Los estudios sociodemográficos que se pretenden realizar con el almacén de datos, están orientados a tomar decisiones gubernamentales para destinar recursos hacia los sectores más necesitados..*"; Y dado que los presupuestos se distribuyen desde los niveles estatales hasta los

niveles municipales, el análisis sociodemográfico buscará obtener información del almacén de datos por diversos niveles sectoriales, lo cual es un factor determinante para el método de particionamiento a utilizar, luego entonces, se tiene que las columnas ENT, MUN y LOC son útiles para un primer particionamiento.

Una vez determinadas estas columnas, se revisa el volumen de datos que se puede tener a nivel entidad federativa, municipio y localidad, además de los distintos valores que se tienen para dichas columnas, y se concluye que la columna ENTIDAD permite un particionamiento adecuado para no tener un gran número de particiones, realizar análisis de información por estado y tener volúmenes manejables por cada una de las 32 particiones que se crearían, dados los 32 estados que conforman la república mexicana.

Una vez determinada la columna a utilizarse para un primer particionamiento, se determina el tipo de particionamiento a utilizar, ya sea RANGOS, LISTA o HASH. En este caso dado que los 32 valores distintos de la columna de particionamiento ENT, distribuyen de manera adecuada los datos entre las particiones, el método HASH queda descartado como opción para el método de particionamiento. También el método de RANGOS, quedaría descartado dado que el tipo de análisis de información que se realizará es por estado, por lo que no tiene sentido establecer "Rangos de Estados" para concentrar en una sola partición varios estados. Por lo tanto para este caso el método de LISTA es el más adecuado a pesar de que la lista que conforma cada partición, sólo está compuesta por una sola entidad o un solo valor, como se observa más adelante en la definición de la tabla.

Por otra parte, el contexto indica que uno de los objetivos que se buscan en la captación de la información del censo es "*Organizar la*

información por municipio durante todo el procesamiento para facilitar su análisis"; por lo tanto se opta también por un SUBPARTICIONAMIENTO por MUNICIPIO. Ahora bien, dada la gran cantidad de municipios que se pueden encontrar y que resulta complejo agrupar los municipios por RANGOS o por LISTAS de municipios, el método adecuado de SUBPARTICIONAMIENTO por MUNICIPIO es el de HASH que realiza una distribución equilibrada y automatizada de las SUBPARTICIONES dentro de cada PARTICION de ESTADO.

Con base en el análisis expuesto en los párrafos anteriores, se tiene la definición de la tabla TR_VIVIENDAS, de la siguiente forma:

```
CREATE TABLE TR_VIVIENDAS
(SECVIV NUMBER,
ENT NUMBER,
MUN NUMBER,
LOC NUMBER,
CLASE_VIV NUMBER ...
...)
PARTITION BY LIST (ent)
SUBPARTITION BY HASH (mun)
Partitions 4
(
PARTITION estado_01 VALUES (1),
PARTITION estado_02 VALUES (2),
PARTITION estado_03 VALUES (3),
PARTITION estado_04 VALUES (4),
PARTITION estado_05 VALUES (5),
PARTITION estado_06 VALUES (6),
....
PARTITION estado_31 VALUES (31),
PARTITION estado_32 VALUES (32)
);
```

Con la instrucción anterior se crea la tabla TR_VIVIENDAS con 32 particiones, cada una de las cuales corresponde a un estado de la República Mexicana que a su vez está subparticionada con 4

subparticiones que corresponderán a algún municipio contenido en el estado asociado a la partición principal.

Análisis de HOGARES

Le estructura de la tabla TR_HOGARES se presenta a continuación:

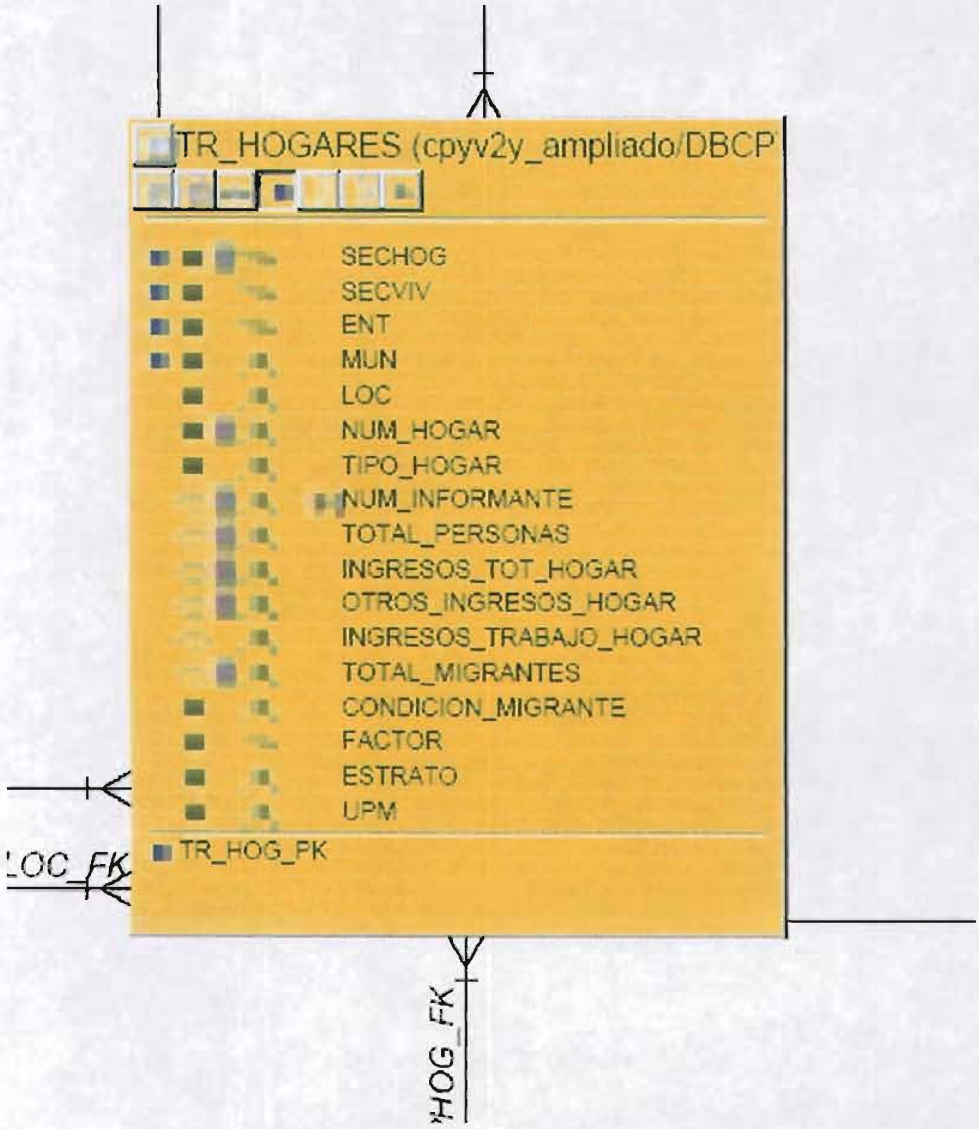


Figura 9 Estructura de la tabla TR_HOGARES

De la estructura anterior, las columnas que son útiles como llaves de particionamiento son:

COLUMNA	DESCRIPCION
SECVIV	Número identificador de la vivienda
ENT	Entidad Federativa donde se ubica el hogar.
MUN	Municipio donde se ubica el hogar.
LOC	Localidad donde se ubica el hogar.

Tabla 2 Columnas que pueden ser llave de particionamiento

Como se puede observar, el análisis realizado para la tabla TR_VIVIENDAS aplica de la misma forma para la tabla TR_HOGARES, ya que la finalidad de esta información es almacenar datos relevantes a la constitución de familias (HOGARES) que pertenecen a una vivienda y dado que estas tablas comparten las columnas de entidad y municipio, las razones expuestas anteriormente permiten concluir que la tabla TR_HOGARES puede ser particionada primeramente utilizando como llave de particionamiento la columna ENT (entidad) para así tener 32 particiones, y de forma secundaria cada una de estas 32 particiones puede subparticionarse por el método HASH utilizando la columna MUN (municipio como llave de subparticionamiento estableciendo 4 subparticiones HASH por cada partición correspondiente a un estado.

Se puede observar en la siguiente instrucción la creación de la tabla TR_HOGARES con los métodos de particionemiento establecidos:

```
CREATE TABLE TR_HOGARES
(SECVIV NUMBER,
ENT NUMBER,
MUN NUMBER,
LOC NUMBER,
TIPO_HOGAR NUMBER,
NUM_HOGAR NUMBER ...
...)
PARTITION BY LIST (ent)
SUBPARTITION BY HASH (mun)
Partitions 4
(
PARTITION estado_01 VALUES (1),
PARTITION estado_02 VALUES (2),
PARTITION estado_03 VALUES (3),
PARTITION estado_04 VALUES (4),
PARTITION estado_05 VALUES (5),
PARTITION estado_06 VALUES (6),
....
PARTITION estado_31 VALUES (31),
PARTITION estado_32 VALUES (32)
);
```

Análisis de POBLADORES

Le estructura de la tabla TR_POBLADORES se presenta a continuación:

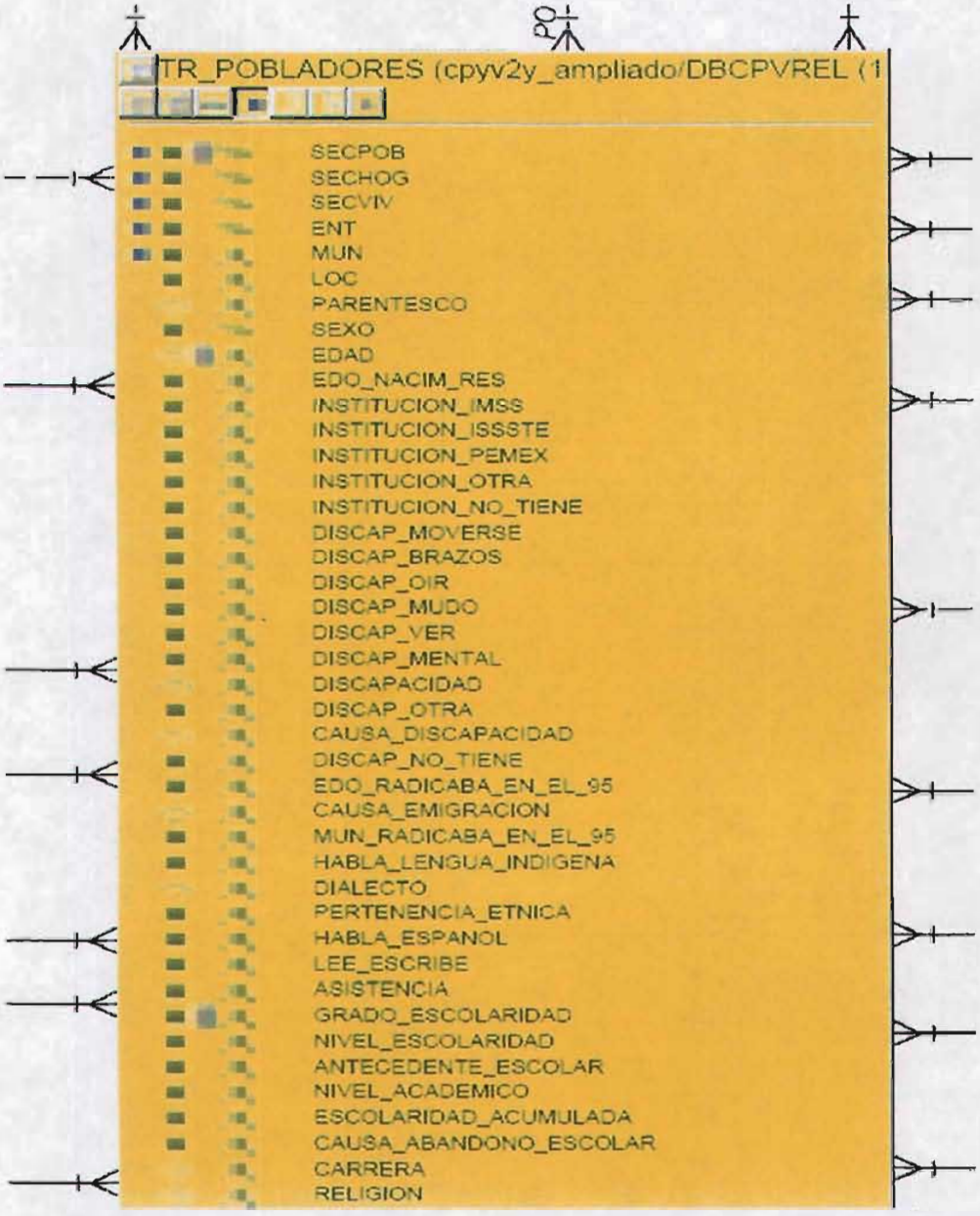


Figura 10 Estructura de la tabla TR_POBLADORES

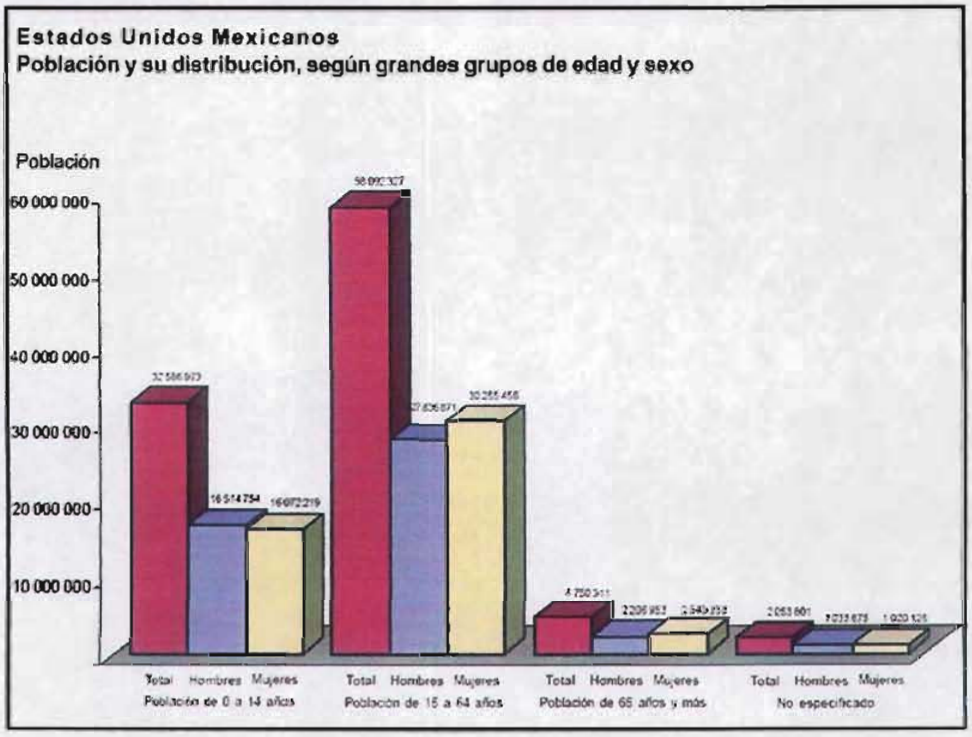
De la estructura anterior, las columnas que son útiles como llaves de particionamiento son:

COLUMNA	DESCRIPCION
SECVIV	Número identificador de la vivienda
ENT	Entidad Federativa donde se ubica el poblador.
MUN	Municipio donde se ubica el poblador.
LOC	Localidad donde se ubica el poblador.
SEXO	Sexo del poblador
EDAD	Edad del poblador

Tabla 3 particionamiento

Para este caso, el análisis realizado en las tablas anteriores aplica de la misma forma para la tabla TR_POBLADORES, es decir, que se tiene un particionamiento por ESTADO y por MUNICIPIO, pero entran en juego 2 columnas que se deben considerar como candidatas a ser llave de particionamiento, las cuales son SEXO y EDAD, ya que se entiende que el análisis de la información censal, está orientado a comprender aspectos socioeconómicos que tienen que ver con el SEXO y EDAD y es de esperarse que el almacén de datos reciba consultas sobre pobladores que caigan en algún rango de edad o que tengan algún género especifico (HOMBRES / MUJERES).

La siguiente gráfica indica el tipo de análisis de la información que se realizará en el almacén de datos:



Gráfica 8 Análisis de la información

La gráfica anterior nos indica la manera de analizar la información por rangos de edades y género; Lo que nos confirma la necesidad de utilizar las columnas EDAD y SEXO como llaves de particionamiento.

Dado lo anterior, la tabla TR_POBLADORES se puede particionar inicialmente por ENTIDAD y cada una de las 32 particiones tendrá 2 subparticiones por SEXO que a su vez serán subparticionadas por rangos de EDAD y finalmente cada partición de EDAD será particionada por MUNICIPIO.

El método de particionamiento por SEXO será POR LISTA, el método de particionamiento por EDAD será por RANGOS y finalmente para particionar por MUNICIPIO se utilizará el método HASH.

La siguiente instrucción muestra la sintaxis para crear la tabla particionada:

```
CREATE TABLE TR_POBLADORES
(SECVIV NUMBER,
ENT NUMBER,
MUN NUMBER,
LOC NUMBER,
SEXO CHAR,
EDAD NUMBER…
...)
PARTITION BY LIST (ent),
SUBPARTITION BY LIST (sexo),
SUBPARTITION BY RANGE (edad)
SUBPARTITION BY HASH (mun)
Partitions 4
(
PARTITION estado_01 VALUES (1),
SUBPARTITION hombres_estado_01 VALUES('H'),
SUBPARTITION hombres_estado_01_edad_0_14 values less than 14,
SUBPARTITION hombres_estado_01_edad_15_64 values less than 64,
SUBPARTITION hombres_estado_01_edad_65_120 values less than 120,
SUBPARTITION hombres_estado_01_edad_nodefinida values null,
SUBPARTITION mujeres_estado_01 VALUES('M'),
SUBPARTITION mujeres_estado_01_edad_0_14 values less than 14,
SUBPARTITION mujeres _estado_01_edad_15_64 values less than 64,
SUBPARTITION mujeres _estado_01_edad_65_120 values less than 120,
SUBPARTITION mujeres _estado_01_edad_nodefinida values null,

PARTITION estado_02 VALUES (2),
SUBPARTITION hombres_estado_02 VALUES('H'),
SUBPARTITION hombres_estado_02_edad_0_14 values less than 14,
SUBPARTITION hombres_estado_02_edad_15_64 values less than 64,
SUBPARTITION hombres_estado_02_edad_65_120 values less than 120,
SUBPARTITION hombres_estado_02_edad_nodefinida values null,
 SUBPARTITION mujeres_estado_02 VALUES('M'),
SUBPARTITION mujeres_estado_02_edad_0_14 values less than 14,
SUBPARTITION mujeres _estado_02_edad_15_64 values less than 64,
SUBPARTITION mujeres _estado_02_edad_65_120 values less than 120,
```

```
SUBPARTITION mujeres _estado_02_edad_nodefinida values null,

PARTITION estado_03 VALUES (3),
SUBPARTITION hombres_estado_03 VALUES('H'),
SUBPARTITION hombres_estado_03_edad_0_14 values less than 14,
SUBPARTITION hombres_estado_03_edad_15_64 values less than 64,
SUBPARTITION hombres_estado_03_edad_65_120 values less than 120,
SUBPARTITION hombres_estado_03_edad_nodefinida values null,
 SUBPARTITION mujeres_estado_03 VALUES('M'),
SUBPARTITION mujeres_estado_03_edad_0_14 values less than 14,
SUBPARTITION mujeres _estado_03_edad_15_64 values less than 64,
SUBPARTITION mujeres _estado_03_edad_65_120 values less than 120,
SUBPARTITION mujeres _estado_03_edad_nodefinida values null,

....
PARTITION estado_32 VALUES (32)
SUBPARTITION hombres_estado_32 VALUES('H'),
SUBPARTITION hombres_estado_32_edad_0_14 values less than 14,
SUBPARTITION hombres_estado_32_edad_15_64 values less than 64,
SUBPARTITION hombres_estado_32_edad_65_120 values less than 120,
SUBPARTITION hombres_estado_32_edad_nodefinida values null,
 SUBPARTITION mujeres_estado_32 VALUES('M'),
SUBPARTITION mujeres_estado_32_edad_0_14 values less than 14,
SUBPARTITION mujeres _estado_32_edad_15_64 values less than 64,
SUBPARTITION mujeres _estado_32_edad_65_120 values less than 120,
SUBPARTITION mujeres_estado_32_edad_nodefinida values null,
);
```

Como se puede observar en la instrucción anterior, se establece el primer particionamiento por estado, después por SEXO y después por rangos de edades en donde el límite inferior de una partición es el límite superior de la partición que antecede, de tal forma que por ejemplo, el limite inferior de la partición hombres_estado_01_edad_15_64, es el límite superior de la partición hombres_estado_01_edad_0_14.

También en todos los ejemplos anteriores del particionamiento de HOGARES, VIVIENDAS y POBLADORES se establece que el particionamiento por municipio se realiza con el método HASH estableciendo 4 subparticiones HASH por cada subpartición o partición según sea el caso.

El caso de estudio seleccionado y el análisis realizado en este capítulo, han permitido poner en práctica los 3 métodos de particionamiento expuestos en el capítulo 2, lo cuál era el objetivo primordial de este trabajo de tesis.

Como resultado de la aplicación de los métodos de particionamiento se explica a continuación un ejemplo que ilustra los beneficios obtenidos:

Ejemplo

Consideremos las siguientes suposiciones:

1) La tabla TR_VIVIENDAS no ha sido particionada.
2) Los bloques de datos que contienen los registros de la tabla son de 8 Kilobytes.
3) El tamaño promedio de cada registro es de 1 Kilobyte.
4) La tabla contiene 30 millones de registros.
5) A las 12:00 hrs. El usuario 1 consulta las viviendas de la entidad 2 (Puebla) y el municipio 5 de dicha entidad.
6) A las 12:00 hrs. El usuario 2 consulta las viviendas de la entidad 32 (Zacatecas) y el municipio 15 de dicha entidad.
7) Los registros correspondientes a la entidad 2 y el municipio 5 de dicha entidad, están contenidos en 250 bloques.
8) Los registros correspondientes a la entidad 32 y el municipio 15 de dicha entidad, están contenidos en 520 bloques.

Dadas las suposiciones anteriores se derivan las siguientes conclusiones:

1) Para almacenar los 30 millones de registros se necesitan 3750 bloques de 8KB.

2) Dado que la tabla TR_VIVIENDAS no ha sido particionada, para encontrar los registros solicitados por el usuario 1 se tendrán que consultar los 3750 bloques de la tabla y obtener los registros que pertenecen a la entidad 2 y municipio 5.
3) Dado que la tabla TR_VIVIENDAS no ha sido particionada, para encontrar los registros solicitados por el usuario 2 se tendrán que consultar los 3750 bloques de la tabla y obtener los registros que pertenecen a la entidad 32 y municipio 15.
4) Dado que el usuario 1 y el usuario 2 realizan la consulta a la misma hora, se observará contención en los discos que contienen los 3750 bloques, es decir que se generará un cuello de botella dado que existen 2 usuarios consultando los mismos bloques al mismo tiempo.
5) Se realizarán 2980 x 2 = 5960 lecturas de bloques innecesarias dado que la información solicitada por los usuarios 1 y 2 esta contenida solamente en 770 bloques.

Si cambiamos la suposición 1 a que la tabla TR_VIVIENDAS fue particionada por el método de LISTA utilizando la columna ENTIDAD y por el método HASH utilizando la columna MUNICIPIO, se tiene como resultado 32 particiones que a su vez contienen 4 subparticiones cada una, con un total de 128 subparticiones.

Supongamos la siguiente distribución de las 128 subparticiones:

Subpartición	**Contenido**
Subpartición 1	Entidad 1 Municipios 1 al 10 de la entidad 1
Subpartición 2	Entidad 1 Municipios 11 al 20 de la entidad 1
Subpartición 3	Entidad 1

	Municipios 21 al 30 de la entidad 1
Subpartición 4	Entidad 1 Municipios 31 al 40 de la entidad 1
Subpartición 5	Entidad 2 Municipios 1 al 10 de la entidad 2 **(250 bloques de acuerdo a la suposición 7)**
Subpartición 6	Entidad 2 Municipios 11 al 20 de la entidad 2
Subpartición 7	Entidad 2 Municipios 21 al 30 de la entidad 2
Subpartición 8	Entidad 2 Municipios 31 al 40 de la entidad 2
....	
Subpartición 125	Entidad 32 Municipios 1 al 10 de la entidad 32
Subpartición 126	Entidad 32 Municipios 11 al 20 de la entidad 32 **(520 bloques de acuerdo a la suposición 8)**
Subpartición 127	Entidad 32 Municipios 21 al 30 de la entidad 32
Subpartición 128	Entidad 32 Municipios 31 al 40 de la entidad 32

Con la tabla TR_VIVIENDAS particionada como se propone tenemos las siguientes conclusiones.

6) Para resolver la consulta del usuario 1 solo será necesario leer la subpartición 5, ya que contiene exactamente los registros solicitados por el usuario, lo cual requiere solamente la lectura de 250 bloques.

7) Para resolver la consulta del usuario 2 solo será necesario leer la subpartición 126, ya que contiene exactamente los registros solicitados por el usuario, lo cual requiere solamente la lectura de 520 bloques.
8) Aunque los 2 usuarios realizan las consultas simultáneamente, no se observará contención en los discos ya que cada usuario consulta bloques distintos.
9) El total de bloques leídos por los dos usuarios es de 770.
10) En comparación con las lecturas realizadas con la tabla no particionada, se tiene un ahorro de 5960 lecturas de bloques no realizadas.

Conclusiones

El presente trabajo realizado me ha permitido derivar un aprendizaje y una serie de conclusiones que presento a continuación.

Al realizar este trabajo he comprendido que nos encontramos en una era en que la información es un elemento imprescindible para el éxito de las organizaciones y que por tal motivo se han desarrollado tecnologías alrededor de ésta que permiten manejarla con mayor facilidad aunque los volúmenes sobrepasen nuestras expectativas.

El deseo incansable del ser humano por conocer más sobre nosotros mismos y controlar nuestro entorno, nos ha llevado a la necesidad de manipular y analizar grandes volúmenes de información con tecnología que se ha desarrollado rápidamente en el transcurso de los últimos 10 años y que es necesario conocer para tomar ventaja de ella y no rezagarnos en el conocimiento.

Tal es el caso del Instituto Nacional de Estadística, Geografía e Informática, que hoy en día utiliza tecnología informática de vanguardia para cumplir con su misión.

Uno de los retos que enfrenta el INEGI para cumplir con su misión es poner a disposición del público en general, la información recolectada por los censos económicos realizados en años pasados; Ya que al abrir el acceso a esta información a cualquier persona, se enfrenta con un problema de capacidad para soportar la diversidad de consultas que podrían realizarse por cientos o miles de personas deseosas de explotar la información censal para la toma de decisiones.

Dado que los recursos son siempre limitados, es necesario conocer y explotar al máximo lo que la tecnología nos ofrece, y en el ámbito de grandes bases de datos que están a disposición de grandes cantidades de usuarios, esto se convierte en un factor crítico de éxito.

Al poner a disposición grandes volúmenes de información, es necesario minimizar el acceso innecesario a datos durante la realización de consultas, ya que los recursos de cómputo pueden saturarse y no satisfacer las necesidades de los usuarios. Los métodos de particionamiento expuestos en este trabajo ayudan a minimizar el acceso innecesario a datos y así liberan recursos de cómputo y los ponen a disposición de otros usuarios con mayor rapidez.

El mantenimiento de la información almacenada en la base de datos puede volverse una tarea compleja si los volúmenes son grandes y no se utilizan los métodos adecuados para organizarla física y lógicamente, lo cual deriva en problemas de disponibilidad de información crítica que se traduce en retrasos en la toma de decisiones importantes.

Otro beneficio que se obtiene al aplicar los métodos de particionamiento expuestos es que los datos son organizados de tal forma que su depuración y mantenimiento se simplifica y se agiliza, poniéndolos disponibles con mayor rapidez.

La transparencia que tienen estos métodos con respecto a la visión conceptual que los usuarios tienen sobre la información, es algo interesante que destacar como aprendizaje derivado de este trabajo. Refiriéndome a que una vez aplicados los métodos de particionamiento expuestos, los usuarios finales no tienen que preocuparse sobre la organización física de los datos, ellos simplemente deben comprender conceptual y lógicamente como

están modelados los datos para tener acceso a ellos y explotarlos, mientras que detrás de esta vista amigable de los datos, éstos se encuentran organizados eficientemente para un acceso óptimo.

A pesar de que el presente trabajo tiene una connotación primordialmente informática, como egresado de la carrera de actuaría he enfrentado situaciones en las cuales las bases aprendidas en la facultad me han ayudado a comprender que en cualquier área de conocimiento, la información es un factor presente que debe saberse modelar, almacenar y explotar para los fines de dicha área de conocimiento.

No cabe duda que en el caso del Instituto Nacional de Geografía, Estadística e Informática los actuarios juegan un papel importante, primordialmente en el área estadística y tienen una estrecha relación con el ámbito informático para cumplir con su labor, por lo que concluyo que la formación informático-actuarial recibida en la Facultad de Ciencias es el mayor beneficio recibido por la universidad para desarrollarme profesionalmente.

Bibliografía

Dataware Housing Guide. US, Oracle Library, 1998.

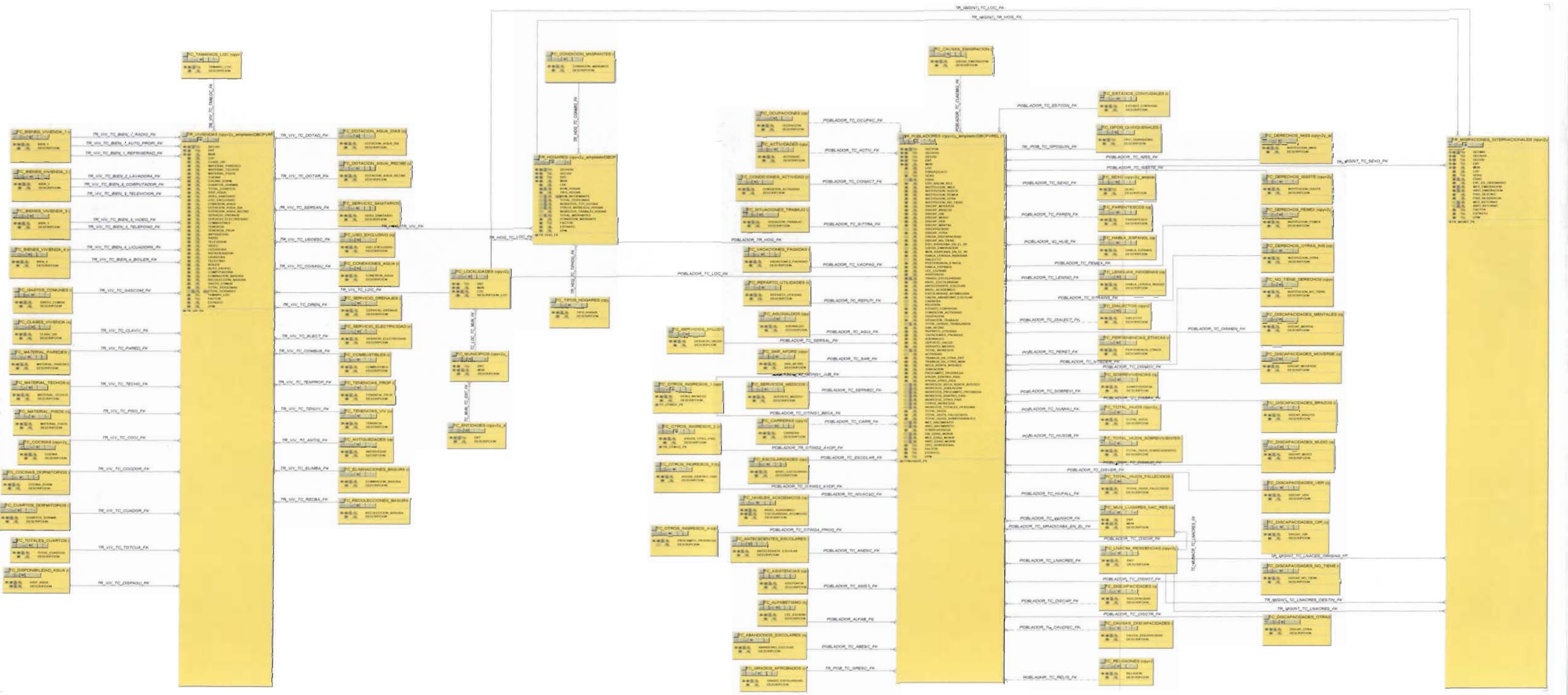
Database Concepts. US, Oracle library, 1995.

Data Warehousing, US, Digital Press, Lilian Hobbs, Susan Hillson, Shilpa Lawande, 2003

Síntesis Metodológica del XII Censo de Población y Vivienda 2000, Instituto Nacional de Geografía Estadística e Informática, México, 2000

Anexo A

Modelo de Datos Censo Nacional de Población y Vivienda 2000


TC_TAMANHOS_LOC (opyv)
TR_VIV_TC_TAMLOC_FK
TC_BIENES_VIVIENDA_1 (o
TR_VIV_TC_BIEN_1_RADIO_FK
TR_VIV_TC_BIEN_1_AUTO_PROP_FK
TR_VIV_TC_BIEN_1_REFRIGERAD_FK
TC_BIENES_VIVIENDA_2 (
TR_VIV_TC_BIEN_2_LAVADORA_FK
TR_VIV_TC_BIEN_2_COMPUTADOR_FK
TR_VIV_TC_BIEN_2_TELEVISION_FK
TC_BIENES_VIVIENDA_3 (
TR_VIV_TC_BIEN_3_VIDEO_FK
TR_VIV_TC_BIEN_3_TELEFONO_FK
TC_BIENES_VIVIENDA_4 (o
TR_VIV_TC_BIEN_4_LICUADORA_FK
TR_VIV_TC_BIEN_4_BOILER_FK
TC_GASTOS_COMUNES (
TR_VIV_TC_GASCOM_FK
TC_CLASES_VIVIENDA (o
TR_VIV_TC_CLAVIV_FK
TC_MATERIAL_PAREDES (
TR_VIV_TC_PARED_FK
TC_MATERIAL_TECHOS (o
TR_VIV_TC_TECHO_FK
TC_MATERIAL_PISOS (o
TR_VIV_TC_PISO_FK
TC_COCINAS (opyv2y_
TR_VIV_TC_COCI_FK
TC_COCINAS_DORMITORIOS
TR_VIV_TC_COCDOR_FK
TC_CUARTOS_DORMITORIOS (
TR_VIV_TC_CUADOR_FK
TC_TOTALES_CUARTOS (
TR_VIV_TC_TOTCUA_FK
TC_DISPONIBILIDAD_AGUA (
TR_VIV_TC_DISPAGU_FK
TR_VIVIENDAS (opyv2y_ampliado)OBCPVIR
TR_VIV_TC_DOTAG_FK
TC_DOTACION_AGUA_DIAS (o
TR_VIV_TC_DOTAR_FK
TC_DOTACION_AGUA_RECIBE (o
TR_VIV_TC_SERSAN_FK
TC_SERVICIO_SANITARIOS
TR_VIV_TC_USOEXC_FK
TC_USO_EXCLUSIVO (o)
TR_VIV_TC_CONAGU_FK
TC_CONEXIONES_AGUA (o
TR_VIV_TC_LOC_FK
TR_VIV_TC_DREN_FK
TC_SERVICIO_DRENAJES (o
TR_VIV_TC_ELECT_FK
TC_SERVICIO_ELECTRICIDAD (o
TR_VIV_TC_COMBUS_FK
TC_COMBUSTIBLES (o
TR_VIV_TC_TENPROP_FK
TC_TENENCIAS_PROP (o
TR_VIV_TC_TENVIV_FK
TC_TENENCIAS_VIV (o)
TR_VIV_TC_ANTIG_FK
TC_ANTIGUEDADES (o
TR_VIV_TC_ELIMBA_FK
TC_ELIMINACIONES_BASURA (o
TR_VIV_TC_RECBA_FK
TC_RECOLECCIONES_BASURA
TC_CONDICION_MIGRANTES (
TR_HOG_TC_CONMIG_FK
TR_HOGARES (opyv2y_ampliado)OBCP
TR_HOG_TC_LOC_FK
TR_HOG_TC_TIPHOG_FK
TC_TIPOS_HOGARES (opy
TC_LOCALIDADES (opyv2y)
TC_LOC_TC_MUN_FK
TC_MUNICIPIOS (opyv2y_
TC_MUN_TC_ENT_FK
TC_ENTIDADES (opyv2y_a
TC_CAUSAS_EMIGRACION (
TR_MIGINT_TC_CAUEMI_FK
TR_MIGINT_TC_LOC_FK
TR_MIGINT_TR_HOG_FK
TR_POBLADORES (opyv2y_ampliado)OBCPVIREL (T
POBLADOR_TC_CAUEMI_FK
TC_OCUPACIONES (op
POBLADOR_TC_OCUPAC_FK
TC_ACTIVIDADES (opy
POBLADOR_TC_ACTIV_FK
TC_CONDICIONES_ACTIVIDAD (o
POBLADOR_TC_CONACT_FK
TC_SITUACIONES_TRABAJO (
POBLADOR_TC_SITTRA_FK
POBLADOR_TR_HOG_FK
TC_VACACIONES_PAGADAS (
POBLADOR_TC_VACPAG_FK
POBLADOR_TC_LOC_FK
TC_REPARTO_UTILIDADES (o
POBLADOR_TC_REPUTI_FK
TC_AGUINALDOS (opy
POBLADOR_TC_AGUI_FK
TC_SERVICIOS_SALUD (
TC_BAR_AFORE (opyv
POBLADOR_TC_BAR_FK
TC_OTROS_INGRESOS_1 (opy
TC_SERVICIOS_MEDICOS (
POBLADOR_TC_SERMED_FK
POBLADOR_TC_OTING1_JUB_FK
TC_OTROS_INGRESOS_2 (o
TC_CARRERAS (opyv2
POBLADOR_TC_CARR_FK
POBLADOR_TR_OTING2_4YOP_FK
TC_OTROS_INGRESOS_3 (o
TC_ESCOLARIDADES (opy
POBLADOR_TC_ESCOLAR_FK
POBLADOR_TC_OTING3_4YOP_FK
TC_NIVELES_ACADEMICOS (op
POBLADOR_TC_NIVACAD_FK
TC_OTROS_INGRESOS_4 (op
POBLADOR_TC_OTING4_PROG_FK
TC_ANTECEDENTES_ESCOLARES
POBLADOR_TC_ANTESC_FK
TC_ASISTENCIAS (op
POBLADOR_TC_ASIS_FK
TC_ALFABETISMO (o
POBLADOR_ALFAB_FK
TC_ABANDONOS_ESCOLARES (o
POBLADOR_TC_ABESC_FK
TC_GRADOS_APROBADOS (o
TR_POB_TC_GRESC_FK
TC_ESTADOS_CONYUGALES (o
POBLADOR_TC_ESTCON_FK
TC_GRUPOS_QUINQUENALES
TR_POB_TC_GPOQUIN_FK
POBLADOR_TC_ISSS_FK
POBLADOR_TC_ISSSTE_FK
TC_SEXO (opyv2y_ampliado
POBLADOR_TC_SEXO_FK
TC_PARENTESCOS (op
POBLADOR_TC_PAREN_FK
TC_HABLA_ESPANOL (op
POBLADOR_VI_HLNE_FK
POBLADOR_TC_FEMEX_FK
TC_LENGUAS_INDIGENAS (op
POBLADOR_TC_LENING_FK
POBLADOR_TC_OTRASL_FK
TC_DIALECTOS (opyv2
POBLADOR_TC_DIALECT_FK
TC_PERTENENCIAS_ETNICAS (
POBLADOR_TC_PERET_FK
POBLADOR_TC_NTENDR_FK
POBLADOR_TC_DISMEN_FK
TC_SOBREVIVENCIAS (op
POBLADOR_TC_SOBREVI_FK
POBLADOR_TC_DISMEV_FK
TC_TOTAL_HIJOS (opyv2y
POBLADOR_TC_NUMHIJ_FK
TC_TOTAL_HIJOS_SOBREVIVIENTES
POBLADOR_TC_HIJSOB_FK
POBLADOR_TC_DISVER_FK
TC_TOTAL_HIJOS_FALLECIDOS (
POBLADOR_TC_HIJFALL_FK
TC_MUS_LUGARES_NAC_RES (op
POBLADOR_TC_MPIOCOABA_EN_EL_FK
POBLADOR_TC_DISOR_FK
TC_LUGNOM_RESIDENCIA5 (opyv2y)
POBLADOR_TC_LNACRES_FK
POBLADOR_TC_DISNOT_FK
TC_DISCAPACIDADES (o
POBLADOR_TC_DISCAP_FK
POBLADOR_TC_DISCTR_FK
TC_CAUSAS_DISCAPACIDADES (
POBLADOR_TC_CAUDISC_FK
TC_RELIGIONES (opyv2
POBLADOR_TC_RELIS_FK
TC_MIGRACIONES_INTERNACIONALES (opyv2y)
TR_MIGINT_TC_SEXO_FK
TC_DERECHOS_IMSS (opyv2y_ad
TC_DERECHOS_ISSSTE (opyv2y)
TC_DERECHOS_PEMEX (opyv2y)
TC_DERECHOS_OTRAS_INS (opy
TC_NO_TIENE_DERECHOS (opyv
TC_DISCAPACIDADES_MENTALES (o
TC_DISCAPACIDADES_MOVERSE (op
TC_DISCAPACIDADES_BRAZOS (o
TC_DISCAPACIDADES_MUDO (op
TC_DISCAPACIDADES_VER (op
TC_DISCAPACIDADES_OIR (op
TR_MIGINT_TC_LNACRES_ORIGINA_FK
TC_DISCAPACIDADES_NO_TIENE (
TR_MIGINT_TC_LNACRES_DESTIN_FK
TR_MIGINT_TC_LNACRES_FK
TC_DISCAPACIDADES_OTRAS (
TC_MIGRACIONES_TC_LNACRES_FK