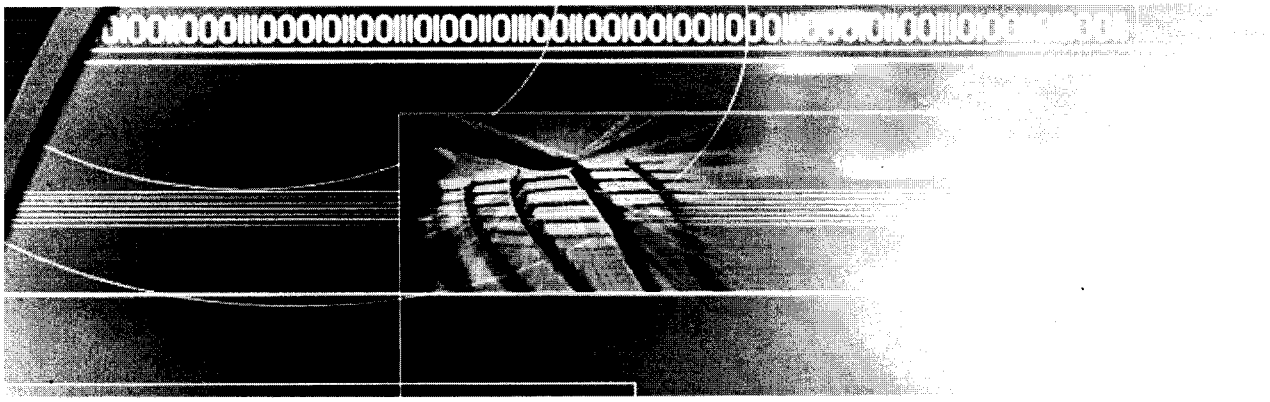
Universidad de las Ciencias Informáticas

Universidad de la Habana

Facultad de Matemática Computación

Carrera Ciencias de la Computación



Servicios Básicos del Ciudadano, Centro de Datos.

Trabajo de Diploma en Opción al Título de

LICENCIADO EN CIENCIAS DE LA COMPUTACIÓN

Autor: Enrique Del Valle Tabares

Tutor: Ing. Alexei Zubizarreta Pérez

CIUDAD DE LA HABANA, 2004

Dedicatoria

A mi familia, y en especial a mis padres...

Agradecimientos

La culminación de una carrera universitaria es uno de los momentos más importantes en la vida de una persona, es un premio al largo período de estudios vencidos.

La preparación de un estudiante universitario es un proyecto en el cual han puesto sus manos muchos profesores, todos con el mismo objetivo de transmitir sus conocimientos y experiencia en la materia impartida. En este gran momento agradezco a todos los profesores que de una forma u otra aportaron su grano de arena a la preparación que he adquirido en los años de estudiante, a los profesores que me ayudaron en la realización de este trabajo y en especial a los profesores de la Universidad de Oriente de la Facultad de Matemática Computación.

A mis padres que siempre tuvieron la seguridad de que podía lograrlo.

A mi hermana.

A mi novia que me apoyó en todos los momentos.

A todos mis compañeros de aula.

A Leonardo Valcárcel que ha cursado conmigo la mayor parte de mi vida de estudiante.

Resumen

En la actualidad todas las esferas de la sociedad se han complejizado por su crecimiento interno y nexos externos, todo ello va acompañado de un gran volumen de datos que deben utilizarse con frecuencia y es entonces que la informática, con sus recursos, entra a jugar un papel importante, especialmente en lo referido a Base de Datos (BD). Siendo una necesidad controlar un gran volumen de información de una forma rápida y eficiente.

Las bases de datos de los grandes sistemas empresariales crecen exponencialmente, siendo cada vez más difícil y costoso realizar algunas operaciones sobre la información, para la solución de este problema se vienen imponiendo los almacenes de datos (data warehousing) y el procesamiento analítico en línea (OLAP), que conforman la base para la toma de decisiones sobre grandes volúmenes de información. Otro problema importante, cuando se habla de grandes volúmenes de información, es la infraestructura del servidor de base de datos, el cual debe contar con las condiciones necesarias para que el sistema sea escalable, el usuario tenga respuestas rápidas a las consultas, presente alta disponibilidad, entre otras características.

En la Universidad de las Ciencias Informáticas se espera que dentro de 5 años tenga una población aproximada de 15 000 habitantes, por lo que las Bases de Datos de los diferentes sistemas aumentarán considerablemente, siendo entonces de suma importancia realizar un estudio profundo de la infraestructura que debe tener el Centro de Datos (Data Center), así como, analizar nuevas tecnologías para la recuperación de grandes volúmenes de información.

Se requiere de la implementación de un algoritmo de búsqueda fonética para que los usuarios puedan encontrar una persona en la BD de la cual no conocen al 100% como se escribe correctamente su nombre. Este tipo de algoritmos sirve para buscar nombres con sonidos similares, brindando un margen de error en la entrada y garantizando en un alto porcentaje una búsqueda exitosa.

Índice

Introducción

Las Bases de Datos es uno de los campos dentro del mundo de la computación al que se le está dedicando estudios intensivos para hacer cada vez más eficiente el acceso a la información.

Las grandes empresas existentes hoy día ya no solo necesitan de la base de datos las transacciones diarias que normalmente se realizan, sino que requieren valores añadidos de esta, como información implícita, análisis estadísticos, etc. Estas nuevas motivaciones han dado origen al Data Warehouse y OLAP (*Online Analytical Processing*), desde el punto de vista de su objetivo final estas tecnologías se basan en una motivación estadística, de obtener medidas sumarias y predicciones en grandes volúmenes de datos. Estas poderosas herramientas permiten a los analistas de las empresas examinar gran cantidad de información y realizar toma de decisiones de una manera eficiente y rápida.

Para la eficiente consulta de los datos no solo basta contar con novedosas técnicas, sino que se necesita de un almacenamiento fiable y que brinde condiciones estables a los clientes. En este punto es donde juegan un papel importante los servidores de BD, requiriendo entre sus características fundamentales una alta disponibilidad, para tener acceso a los datos en cualquier momento.

La UCI será dentro de pocos años una pequeña ciudad de alrededor de 15000 personas, siendo el servicio de buscar una persona en la BD de gran importancia. A los algoritmos de búsqueda de palabras similares se les ha dedicado numerosos estudios y existen varias técnicas para la realización de estos en dependencia de que queremos en nuestro sondeo.

La Universidad de las Ciencias Informáticas necesita de la aplicación de estas novedosas tecnologías, tanto para el análisis de la información como para el almacenamiento confiable de la misma.

Los objetivos del trabajo son:

- Realizar un estudio sobre las herramientas OLAP.

- Hacer una propuesta de infraestructura escalable y de alta disponibilidad para el o los servidores de BD.
- Realizar estudio e implementación de un algoritmo de búsqueda fonética para la exploración de personas.

El trabajo está estructurado en cuatro capítulos:

- Capítulo 1. Fundamentación Teórica: Explica brevemente y brinda una panorámica de los objetivos a cumplir en el proyecto.
- Capítulo 2. Algoritmo de búsqueda fonética: Realiza un estudio sobre diferentes algoritmos para encontrar palabras similares, y se desarrolla un método de búsqueda fonética para el idioma español.
- Capítulo 3. Centro de Datos: Realiza estudio y propuesta de infraestructura para el centro de datos. Basado fundamentalmente en las características de los servidores de BD para que brinden una alta disponibilidad y escalabilidad.
- Capítulo 4. Herramientas OLAP: Explica los aspectos básicos de estas herramientas y muestra las características particulares de Analysis Services de SQL Server 2000.

Capítulo 1

Fundamentación Teórica

1.1 Introducción.

El desarrollo en los últimos tiempos de las Tecnologías de la Informática y las Comunicaciones ha sido motor impulsor en la informatización de la sociedad, y las Bases y Almacenes de Datos tienen un rol fundamental en este desarrollo.

La tecnología de Bases de Datos (BD) no es algo nuevo, se está aplicando desde hace mucho tiempo a la solución de problemas, pero en la actualidad, gran parte de la gestión informática en las empresas y organizaciones modernas gira en torno a estas. La mayoría de las aplicaciones deben almacenar, recuperar y operar con datos, siendo este un punto crítico a tener en cuenta a la hora de diseñar un sistema.

Los datos cada día adquieren más importancia en la toma de decisiones y descubrimientos de conocimientos, ampliando las funcionalidades que se demandan de las BD, siendo este campo, a pesar de ser tan viejo, objeto de un profundo estudio y amplio campo de investigación.

1.2 Bases de Datos.

En los inicios de las BD las estructuras de datos eran muy simples y el costo para procesar y acceder a la información era muy alto. A la par con el desarrollo de la tecnología han evolucionado los procedimientos y estructuras para el almacenamiento y procesamiento de la información. Con el tiempo y las nuevas investigaciones sobre la tecnología de BD se tuvo la necesidad de independizar la estructura de la información de los procedimientos, dando lugar al concepto de Base de Datos: conjunto de datos persistentes que es utilizado por los sistemas de aplicación de alguna empresa dada. [1]. En este contexto el término empresa

se refiere a cualquier organización, lo cual incluye la posibilidad de que sea un solo individuo que es usuario de la BD.

"¿Cuánta información hay en el mundo?" De acuerdo a Michael Lesk, la enorme cantidad de datos tomaría varios miles de millones de gigabytes o varios miles de petabytes de almacenamiento [2]. La tecnología de las modernas bases de datos provee los medios para almacenar, administrar y acceder a semejantes cantidades de información. Sin embargo, las bases de datos no son un fenómeno nuevo. De hecho, Herman Hollerith mecanizó el almacenamiento del Censo de EE.UU. de 1890 en lo que de alguna forma se considera la primera base de datos significativa "computarizada" [3].

Desde hace mucho tiempo la mayoría de la información de todo el mundo está contenida en bases de datos, de aquí su merecida importancia y desarrollo vertiginoso para poder estar a la par con las necesidades de administración y transacciones diarias. Por estas razones los administradores y diseñadores de bases de datos enfrentan numerosos retos para expresar y mantener el complejo ambiente de los datos, sus relaciones, así como la seguridad, integridad y recuperación rápida y fiable de estos.

1.2.1 Rendimiento y escalabilidad.

Cuando se va a realizar una aplicación de BD se tiene que pensar en todas las partes que conforman la misma, desde el diseño lógico de los datos, hasta el gestor de BD más adecuado y la arquitectura del servidor que vamos a disponer.

En este tipo de aplicaciones son muy importantes los términos de rendimiento y escalabilidad. El rendimiento se tiene que tener en cuenta desde el principio, y algo que influye mucho en el mismo es un buen diseño de la BD. El rendimiento es una medida del número de transacciones simultáneas que puede controlar la base de datos correctamente, y para que una aplicación de BD sea escalable debe admitir el rendimiento requerido para el número previsto de usuarios simultáneos.

Estos términos se pueden mejorar de muchas maneras según las características de nuestro sistema, y después de realizar un profundo estudio de muchos puntos tanto del diseño de la BD, como de software y hardware.

Los sistemas para mejorar la escalabilidad pueden crecer de dos formas fundamentales:

- **Escalado vertical:** se logra agregando más hardware a un único nodo o actualizando a un nodo más grande.
- **Escalado horizontal:** se logra agregando más nodos y distribuyendo los datos para repartir la carga de trabajo entre ellos.

Los términos rendimiento y escalabilidad están entre las principales preocupaciones de los diseñadores de sistemas que manejan gran volumen de información. A la par que las empresas crecen y concentran más información se hace necesario buscar una solución para los altos niveles de transacción diarias. Con cada aumento se presentan nuevos retos de rendimiento y escalabilidad.

1.3 Tendencias actuales

Las bases de datos y su tecnología han tenido y están teniendo un impacto considerable y decisivo en el mundo de la informática, siendo una necesidad manejar grandes volúmenes de información de forma rápida, ágil y segura.

Los datos encaminados cada vez más al enfoque comercial y de reglas de negocios, están cambiando su forma de almacenamiento y el uso que se hace de ellos, siendo de especial interés la extracción de información implícita en la BD, surgiendo la necesidad de plantear nuevos modelos y sistemas de BD que aporten un valor añadido a las BD relacionales.

Esto abre nuevos campos dentro de la Tecnología de BD, como son los Almacenes de Datos (Data Warehouse), Exploración de los Datos (Data Mining), y OLAP (*Online Analytical Processing*) para la recuperación de grandes volúmenes de información.

La información está conformada por los datos y los significados de los mismos, el dato por sí solo no es información y sobre esta base trabajan los AD que constituyen la base del proceso analítico (OLAP), para brindar la información necesaria en el momento en el momento pedido. A diferencia de las BD, los AD, contienen datos consolidados, rico contenido histórico, estructura de datos comprensibles y son eficientes en las consultas. Los datos contenidos en un almacén de datos se encuentran organizados para permitir el análisis más que para procesar transacciones en tiempo real como ocurre en los sistemas de proceso de transacciones en línea (OLTP, *Online Transaction Processing*).

1.4 OLAP (*Online Analytical Processing*)

La tecnología OLAP permite un uso más eficaz de los almacenes de datos para el análisis en línea, lo que proporciona respuestas rápidas a consultas analíticas complejas e iterativas. Los modelos de datos multidimensionales de OLAP y las técnicas de datos agregados organizan y resumen grandes cantidades de información para que puedan ser evaluados con rapidez mediante el análisis en línea y las herramientas gráficas. La respuesta a una consulta realizada sobre datos históricos a menudo suele conducir a consultas posteriores en las que el analista busca respuestas más concretas o explora posibilidades. Los sistemas OLAP proporcionan la velocidad y la flexibilidad necesarias para dar apoyo al analista en tiempo real. Mientras que los almacenes y puestos de datos es donde se guardan los datos, el Proceso Analítico en Línea (OLAP) es la tecnología que permite a las aplicaciones de cliente el acceso eficiente a los mismos.

Para mejorar la eficiencia de las consultas analíticas, OLAP proporciona una representación multidimensional de los datos mediante la creación de cubos, los que pueden diseñarse de una forma fácil o difícil según la estructura del almacén de datos.

OLAP y los Almacenes de Datos (AD) son tecnologías que no se pueden separar una de otra, por ser estos la base para el proceso analítico en línea. Los AD contienen datos históricos, muchas veces recopilados de varios orígenes, por lo que son los encargados de combinar estos datos, unificarlos, limpiarlos y

organizarlos para facilitar la recuperación de la información. Por estas razones el AD es una herramienta poderosa para la toma de decisiones de cualquier organización, porque contiene la historia de los procesos de la misma, siendo de vital importancia su consulta a la hora de ejecutar nuevos proyectos.

"Un Data Warehouse o Depósito de Datos es una colección de datos orientado a temas, integrado, no volátil, de tiempo variante, que se usa para el soporte del proceso de toma de decisiones gerenciales". [4]

La creación de un AD es una tarea compleja, se deben tener en cuenta muchas consideraciones debido a que el objetivo de un AD es muy distinto al de un sistema de transacción de datos (OLTP), del primero se espera la organización de un gran volumen de información para su posterior recuperación y análisis, mientras que en el segundo se espera una gran velocidad en las transacciones de modificación, inserción, etc.

Según estudios realizados el número de profesionales que han planeado crear un almacén de datos ha aumentado considerablemente con respecto a años anteriores, las ventajas que brindan los AD y OLAP son fácilmente visibles para cualquier empresa en la que se requiera un diseño completamente orientado al negocio y a la toma de decisiones.

1.5 Algoritmos de búsqueda fonética

El problema de buscar información en Bases de Datos es bastante viejo, y junto con él el problema de buscar palabras similares. Para este tipo de búsqueda existen varios algoritmos, por ejemplo: la diferencia entre dos palabras, el número de la mayor subcadena común entre las dos palabras, de Hamming Distance, etc.

Estos algoritmos resuelven en gran medida el problema de buscar palabras similares según la escritura, pero no según la pronunciación. La búsqueda fonética no es algo nuevo, desde hace mucho se vio la necesidad de buscar datos de personas de las cuales no conocíamos exactamente la ortografía de su nombre pero sí la pronunciación. Uno de los algoritmos más viejos creado para la solución de este problema es el Soundex, el algoritmo fue inventado por **Donald**

Knuth pero el método original fue desarrollado por **Margaret K. Odell y Robert C. Russell.** Más adelante se desarrollo otro método muy conocido llamado Metaphone, que es una modificación del método Soundex.

Estos algoritmos son completamente dependientes del lenguaje, debido a que la fonética de cada idioma es distinta y las reglas varían de uno a otro. Debido a este problema para analizar y emprender la realización de un método de este tipo para una lengua específica hay que hacer un estudio profundo del lenguaje y de sus particularidades.

En general estos métodos convierten cada palabra a un código fonético, y si dos palabras presentan el mismo código es porque presentan similitud fonética. Esto se realiza agrupando las letras según su similitud de pronunciación y dándole un código representante a cada grupo, con esta tabla y ciertas reglas particulares del idioma se transforma cualquier palabra al código fonético.

La aplicación de métodos de este tipo es inmensa y de alta demanda, brindándole al usuario un margen de error en la entrada y garantizándole el éxito de la búsqueda.

1.6 Resumen

En resumen podemos plantear que la tecnología de bases de datos es algo que se desarrolla a la par de la sociedad y sus necesidades. Los sistemas actuales manejan un gran volumen de información, siendo la historia de las transacciones de la empresa u organización.

Teniendo en cuenta el crecimiento anual tanto de los usuarios como de la información requerida en la UCI, los temas planteados en este capítulo son de vital importancia para el buen funcionamiento del sistema integrado que se quiere lograr en un futuro. La BD ***Persona*** por ejemplo forma parte del núcleo de muchos sistemas, demandando gran importancia su disponibilidad, capacidad de respuesta desde disímiles lugares del sistema y bajo cualquier circunstancia.

Muchas empresas están migrando poco a poco sus sistemas de BD a AD para lograr de esta forma una mejor estructura para la toma de decisiones mediante la tecnología OLAP, la que se impone poco a poco mostrando sus ventajas en las complejas consultas analíticas. La UCI necesita de esta tecnología en varios sistemas que deben entregar reportes analizando un alto de grado de información.

Capítulo 2

Algoritmo de búsqueda fonética

2.1 Introducción.

En este capítulo realizaremos un breve estudio de los algoritmos de búsqueda de palabras similares y el desarrollo de uno de ellos, la búsqueda fonética, en la que se encuentran palabras fonéticamente iguales o sea palabras que pueden ser alfabéticamente diferentes pero con similar pronunciación.

La UCI en un futuro se espera que contenga un aproximado de 15 000 personas en su sistema, tanto trabajadores como estudiantes, siendo la encuesta a datos de las personas algo común y diario por muchas aplicaciones para diversos fines, de esto deriva la importancia del método planteado al realizar una búsqueda en la que no conocemos al 100% el nombre de la persona deseada.

2.2 Introducción a los algoritmos de búsqueda de palabras similares.

La búsqueda de palabras similares es un tema al que se le ha dedicado tiempo e investigación. Estos métodos son muy utilizados, principalmente en la búsqueda de bases de datos. Existen diferentes algoritmos, según el tipo de característica similar que se esté buscando y lo que deseamos que el mismo devuelva. Existen dos clases fundamentales de algoritmos para buscar cadenas similares, los métodos equivalentes y los métodos de ranking de similitud.

Los métodos de equivalencia comparan dos cadenas y devuelven por lo general verdadero o falso, según las reglas de equivalencia que se hayan implementado en el algoritmo. Mientras que los métodos de ranking de similitud comparan una cadena con otra y lo que devuelven es un número al

que le podemos llamar como la distancia que existe entre las dos palabras, o sea, dado un grupo de palabras será más similar a la palabra dada la que menor distancia tenga.

Breve explicación sobre algunos algoritmos.

De equivalencia:

Word Stemming: este método reduce la palabra a su forma canónica, o sea a la forma más simple. Por ejemplo dada la palabra *nadaron,* el método la transforma a nadar. Al comparar dos palabras lo primero que se realiza es llevar ambas a su forma simple y compararlas. Como se observa fácilmente es un método idioma-dependiente, se necesita para su implementación un diccionario básico de formas simples y se ha demostrado que pueden haber problemas donde existen diferencias regionales en el idioma.

Expresiones Regulares: este método es muy utilizado en la búsqueda de nombres de ficheros en las computadoras, por ejemplo para buscar todos los archivos que contengan la extensión "jpg", ponemos "*.jpg", en este caso el patrón "*" significa cualquier combinación de caracteres o ninguno. En general estos algoritmos se basan en un grupo de patrones con distintos significados que se combinan con la subcadena que queremos se encuentre en todas las palabras encontradas. Esto brinda gran facilidad al usuario para encontrar lo que realmente desea construyendo la expresión regular óptima para la búsqueda.

Búsqueda Fonética: los algoritmos de este tipo encuentran palabras con sonidos similares. En general estos llevan las palabras a un código fonético, dos cadenas con el mismo código poseen similar pronunciación. Fácilmente se observa que es idioma-dependiente, según el idioma las consonantes se agrupan en diferentes llaves según su similitud fonética.

De ranking de similitud:

Subcadena Común más Larga (Longest Common Substring): este método devuelve un número que no es más que la cantidad de caracteres de la subcadena común más larga que existe entre dos palabras. Ej. Si comparamos

"***computa**ción*" y "***computa**dora*" el método devuelve 7, que es el tamaño de la subcadena ***computa.*** Una desventaja es que la posición del error afecta completamente la similitud de las palabras. Si el error ocurre en el medio divide en dos una posible subcadena más larga, a diferencia si el error estuviera al principio o al final.

Distancia de Edición: este método se basa en los errores principales de mecanografía, como son omisión de caracteres, inserción, substituciones e inversiones. Dada dos palabras devuelve la cantidad de operaciones a realizar para convertir una cadena en otra. Si la distancia es 0 es porque las palabras son idénticas.

Hamming Distance (Distancia de Hamming): devuelve el número de posiciones en el cual los caracteres de las palabras son diferentes. En la comparación de dos cadenas de diferentes tamaños se pueden tomar dos consideraciones: devolver que la distancia es infinito o completar con un carácter especial los espacios a la derecha de la menor cadena.

Muchos sistemas le brindan al usuario la recuperación de información a través de una búsqueda a partir de una cadena clave. La habilidad del sistema para encontrar la información pertinente basada en la entrada del usuario es importante a un sistema exitoso. Para escoger uno de los métodos debemos analizar que es lo que queremos en realidad. ¿Qué tipo de igualdad se quiere reconocer?, ¿la búsqueda es sobre un mismo idioma?, ¿cuan rápido debe ser el algoritmo?, etc. Muchos sistemas utilizan un hibrido entre los dos tipos de algoritmos, para implementar un método de dos pasos, donde en el primer paso se recupera información utilizando un algoritmo de equivalencia, y después se refina esta búsqueda con un método de ranking de similitud.

2.3 Soundex.

Margaret K. Odell y Robert C. Russell fueron los desarrolladores de este algoritmo, basados en el problema de buscar nombres de personas en diversos sistemas como por ejemplo, el sistema de censos. Muchas veces al realizar una búsqueda se conoce la palabra según la pronunciación pero no sé

sabe como se escribe correctamente, el algoritmo Soundex resuelve parcialmente estos tipos de problemas basándose en la similitud fonética más que en la ortografía.

"*La fonética es una rama de la lingüística que estudia la sustancia de los sonidos; la materia fónica y la capacidad que tiene para asociarse con significados específicos: cómo se pronuncian las letras, que características acústicas poseen y cómo se diferencian*". [FON01]

En general el método Soundex reúne las consonantes según su similitud fonética en grupos, y siguiendo ciertas reglas convierte cada palabra al código soundex. Por esta característica el algoritmo es dependiente del idioma y se necesitan realizar cambios para su utilización sobre otra lengua. El método Soundex original fue desarrollado para el idioma Ingles.

2.3.1 Reglas del algoritmo para el idioma ingles.

El algoritmo SOUNDEX lleva cada consonante de la palabra al código que representa el grupo al que pertenece, como puede ser un carácter.

Los grupos para el idioma inglés son los siguientes:

- **1 : B, F, P, V**
- **2 : C, G, J, K, Q, S, X, Z**
- **3 : D, T**
- **4 : L**
- **5 : M, N**
- **6 : R**

El método devuelve un código compuesto de cuatro caracteres, donde el primero es la primera letra de la palabra, y los tres restantes representan los grupos de las tres primeras consonantes diferentes. El método sigue las siguientes reglas:

- Las consonantes Y, W y H no se toman en cuenta.

- Las vocales se desprecian, excepto que una vocal sea la primera letra de la palabra.

- Si dos o más consonantes consecutivas pertenecen al mismo grupo solamente se tiene en cuenta a la primera.

- Sí la primera y segunda letras son consonantes y la segunda pertenece al mismo grupo de la primera, la segunda se desprecia.

2.4 Metaphone.

El algoritmo Metaphone se basa en el método Soundex pero incluye numerosas condiciones adicionales basadas en modelar de una forma más fuerte la fonética del idioma Inglés. Este método mejora los grupos encontrados en comparación con el Soundex, acercándose más al resultado final que el usuario desea.

Explicación del algoritmo.

Elimina las vocales excepto si está en la primera posición.

Si existen letras repetidas quita la segunda excepto si es "cc".

Convierte el alfabeto a 16 grupos de sonidos (B X S K J T F H L M N P R 0 W Y) siguiendo ciertas transformaciones, veamos algunas de ellas:

- C → X si -cia- or -ch-
  S si -ci-, -ce- or -cy-
  K en otro caso incluyendo -sch-
- D → J si -dge-, -dgy- or -dgi-
  T en otro caso.
- L → L
- M → M
- N → N
- T → X si -tia- or -tio-
  0 si antes está "h"
  no suena si está en -tch-
  T en otro caso.

Consideraciones en el inicio de la palabra:

- kn-, gn- pn, ac- o wr- elimina la primera letra.
- x cambia por s.
- wh cambia por w.

Una de las características fundamentales que posee es que no incluye la primera letra dentro del código que genera, lo cual traía problemas en ciertas palabras que eran fonéticamente iguales y que el Soundex no las reconocía, por ejemplo:

***Cline y Kline*:** estas palabras producen un código soundex diferente, sin embargo son fonéticamente iguales.

Este nuevo algoritmo realmente revoluciona con sus condiciones al método anterior, teniendo una fuerte lista de transformaciones basadas en la sonoridad del idioma inglés.

Un algoritmo de este tipo se puede realizar tan fuerte como se quiera incrementando las condiciones del idioma en mayor o menor grado, lo cual se deja a decisión del creador según la utilización que se le va a dar al mismo.

2.5 Desarrollo del algoritmo para el idioma español.

Después de analizar los principales algoritmos existentes para resolver este tipo de búsquedas, se demuestra, que lo mejor es tener aparte de los grupos de equivalencia fonética del idioma, reglas que tengan en cuenta las combinaciones de letras que hacen que una consonante cambie su sonido o realicen un sonido específico.

Los grupos de equivalencia fonética para el idioma español son los siguientes:

- **1- S, C, Z**
- **2- V, B**
- **3- LL, Y**

- **4- G, J, X**
- **5- K, Q, C**
- **6- P**
- **7- D**
- **8- R**
- **9- L**
- **B- M, N**
- **C- F**
- **D- T**
- **F- Ñ**
- **G- W**

La tabla de equivalencias fonéticas fue obtenida de una tesis de la **Universidad de las Américas-Puebla,** Escuela de Ingeniería, Departamento de Ingeniería en Sistemas Computacionales presentada por Patricia García Jiménez. El trabajo consiste en desarrollar un sistema para almacenar y consultar libros en una biblioteca digital [5].

Transformaciones generales:

- **H** → "" (es silente).
- **Y** → **LL** si está seguida de vocal.

  "" en otro caso (se elimina porque esta haciendo función de vocal).
- **G** → **W** si está seguida de **ua** o **uo**.
- **C** → **K** si **ca, co, cu, cr, cl.**

  S si **ce, ci.**

  K en otro caso.

Transformaciones en el inicio de la palabra:

- **U** → **W** si **ua, uo, ue.**
- **hie** → **ye**.

- **io → yo.**
- **X → S.**

Después de realizarle estas transformaciones a la palabra encuestada se eliminan todas las vocales, excepto el caso que la vocal esté en la primera posición de la palabra, y se busca el grupo de las primeras 4 consonantes. Esta opción de las primeras cuatro consonantes puede ser configurable para hacer la búsqueda más específica aumentando el número de consonantes a procesar. A diferencia del algoritmo Soundex que incluye dentro del código la primera letra este algoritmo para el idioma español solo incluye la primera letra si es vocal sino se convierte también a código. En caso de que la palabra no tenga cuatro consonantes distintas se completa con ceros.

2.6 Resumen.

En el estudio realizado se han tenido en cuenta transformaciones básicas, dado que el algoritmo se utilizará para la búsqueda de nombres similares y entonces ciertas combinaciones en los nombres propios son difíciles de encontrar. Este método brinda una posibilidad diferente para encontrar un usuario deseado o un conjunto de ellos.

Para lograr que la comparación con los registros de la BD sea lo más rápida posible, se propone crear un campo en la tabla donde se encuentre el nombre de las personas para almacenar el código correspondiente devuelto por el algoritmo, esta es una opción que mejora en gran medida la rapidez de la búsqueda, reduciéndose a una simple comparación de cadena. Este código se insertaría en el mismo momento en que se insertan los otros datos de la persona, por lo que no constituye ningún conflicto para la integridad de la BD.

Capítulo 3

Centro de Datos

3.1-Introducción.

La tecnología de BD ha mostrado un avance vertiginoso en los últimos años, han aparecido manejadores de BD más potentes y eficientes, nuevas formas de organizar y almacenar la información, como los Data Warehouse, y técnicas eficientes de recuperación de datos como la tecnología OLAP, pero el avance de la tecnología va acompañado del desarrollo del hardware y del uso eficaz del mismo para lograr aplicaciones competentes.

A medida que aumenta el número de usuarios, las aplicaciones deben hacer frente a elevadísimas cargas de conexiones y transacciones. En respuesta a esta necesidad se necesita instalar un potente servidor de BD que pueda administrar miles de conexiones y terabytes de información a la vez de mantener una alta disponibilidad. Los sistemas de equipos escalables pueden incrementar la base cliente, la base de datos y el rendimiento de las aplicaciones sin necesidad de reprogramación.

En la UCI se necesita de un centro de datos que posea una infraestructura escalable y de alta disponibilidad que permita de forma transparente a los usuarios actualizarse a medida que aumentan las BD y el número de conexiones posibles. En este capítulo nos centraremos en mostrar una solución de arquitectura para el servidor de BD del Data Center.

3.2- Sistemas Gestores de BD (SGBD).

Los Sistemas Gestores de Bases de Datos (SGBD) son una parte importante dentro del mundo de la información, contienen todas las rutinas necesarias para el manejo de los datos; podemos definir un SGBD como: "*conjunto de herramientas que suministra a todos (administrador, analistas, programadores, usuarios) los medios necesarios para describir, recuperar y manipular los datos almacenados*

en la BD, manteniendo la seguridad, integridad y confidencialidad de los mismo". [6]

Existen actualmente numerosos SGBD, como por ejemplo: Access, Oracle, SQL, PostgreSQL, MySQL, etc., cada sistema presenta características propias, la elección de uno u otro sistema para gestionar nuestros datos vendrá definida por nuestras necesidades.

Objetivos de los SGBD.

Los SGBD permiten un control total de la información, los principales objetivos son:

- Evitar la redundancia de los datos, eliminando así la inconsistencia de los mismos.
- Mejorar los mecanismos de seguridad de los datos y la privacidad.
- Asegurar la independencia de los programas y los datos, es decir, la posibilidad de modificar la estructura de la base de datos (esquema) sin necesidad de modificar los programas de las aplicaciones que manejan esos datos.
- Mantener la integridad de los datos realizando las validaciones necesarias cuando se realicen modificaciones en la base de datos.
- Mejorar la eficacia de acceso a los datos, en especial en el caso de consultas imprevistas.

Funciones de los SGBD.

Las principales funciones que debe realizar un SGBD son:

- La definición de los datos.
- La manipulación de los datos.
- Garantizar la seguridad e integridad de los datos.
- La gestión de las transacciones y el acceso concurrente.

Actualmente existen dos SGBD líderes del mercado mundial SQL Server y Oracle, vamos a ver algunas características de ambos.

3.2.1 SQL Server 2000.

Entre los principales SGBD se encuentra SQL Server 2000, el cual ha demostrado por sus características y pruebas realizadas un fuerte candidato a tener en cuenta.

Algunas características de SQL Server 2000 Enterprise que lo convierten en un potente gestor de BD [7]:

- Utiliza las características de Windows 2000 Server y Windows Server 2003 para crear los grandes servidores.
- Utiliza las ventajas que brinda el hardware haciendo uso de múltiples procesadores para poder ejecutar más subprocesos
- Usa de forma eficiente la memoria principal almacenando la mayor parte posible de información relativa a la BD.
- El motor relacional admite el procesamiento de transacciones de gran velocidad y el uso de las aplicaciones de almacén de datos más exigentes.
- El motor de ejecución de consultas utiliza los sistemas multiprocesador y multidisco a través de combinaciones hash híbridas paralelas y fusiones de combinaciones.
- El ejecutor de consultas utiliza memorias principales muy grandes, una E/S asíncrona de gran capacidad y paralelismo interno para obtener un mejor rendimiento SMP en las consultas de toma de decisiones.
- Admite las vistas indexadas que son esenciales para las aplicaciones orientadas a informes.
- Trae incorporada la herramienta Analysis Services, para la creación y procesamiento de cubos de datos, basado en la Tecnología OLAP.

SQL Server presenta numerosas ventajas además de ser un SGBD de fácil utilización, en nuestro estudio nos centraremos en la capacidad de escalabilidad y disponibilidad del sistema.

Escalabilidad y disponibilidad.

A medida que las organizaciones crecen y tienen en su poder más datos, es preciso buscar una solución para las grandes cargas de trabajo de las transacciones y las bases de datos de gran tamaño. Con cada incremento se presentan nuevos retos de escalabilidad.

En general un sistema se considera escalable si puede responder correctamente a los siguientes crecimientos:

- Crecimiento de la Base de Datos.
- Crecimiento del número de usuarios y con ello del número de transacciones.
- Crecimiento de la complejidad de las transacciones.
- Crecimiento del número de operaciones que hace uso de la BD.

Un sistema escalable evita la necesidad de crear un sistema completamente nuevo con nuevos componentes de software y hardware cada vez que el sistema alcanza sus límites de capacidad. El sistema sigue ejecutando el mismo software y solo se necesitan agregar recursos de hardware según lo requerido para soportar el aumento de actividad. Las aplicaciones no necesitan recrearse o rediseñarse cuando se agregan recursos.

SQL Server 2000 cuenta con características de escalabilidad y confiabilidad de elevadas prestaciones, entre las que destacan [7]:

- Trasvase de registros para servidores secundarios de reserva.
- Vistas de particiones actualizables entre nodos de clústeres.
- Posibilidad de trabajar con memoria de gran tamaño.
- Posibilidad de trabajar con SMP (hasta 64 procesadores).

- Posibilidad de trabajar con grandes clústeres de Windows Server 2003 Datacenter Server.
- Posibilidad de trabajar con varias instancias de SQL Server 2000 en un solo servidor.
- Integración con Active Directory para ofrecer a los servidores que ejecutan SQL Server acceso transparente a la ubicación.
- Paralelismo mejorado en las operaciones de administración de datos y bases de datos.
- Vistas indexadas y esquema de copos de nieve (snowflake) para admitir almacenes de datos a gran escala.
- Alta disponibilidad: compatibilidad con clústeres de conmutación por error de Windows NT, Windows 2000 y Windows Server 2003.
- Posibilidad de distribuir la carga de trabajo entre varios servidores, diseñando servidores de datos federados (unión de servidores).
- Compatibilidad XML nativa para operaciones de Internet y de intercambio de datos.
- Servicios de notificación para permitir al almacenamiento en la caché cliente y las aplicaciones de mensajería.

Microsoft® SQL Server™ ha evolucionado para adaptarse a las bases de datos y aplicaciones de grandes dimensiones, incluidas las bases de datos de varios terabytes de capacidad compartidas por miles de personas. Logra la escalabilidad soportando los dos tipos principales de crecimiento existentes como son el escalado vertical y horizontal.

-Escalado vertical.

El escalado vertical se logra agregando más potencia al hardware de un único nodo o actualizando a un nodo más grande. O sea con multiprocesadores simétricos (SMP), agregando más procesadores, memoria, discos y tarjetas de red a un solo servidor. La arquitectura de software utilizado por SMP es llamada

modelo de memoria compartida, ejecuta una copia del sistema operativo con los procesos de aplicación como si éstos se encontraran en un sistema monoprocesador. SQL Server 2000 está diseñado para disponer de escalabilidad en sistemas SMP.

Los límites prácticos actuales para el uso general en un solo nodo SMP son:

- 64 procesadores.
- 512 gigabytes de memoria principal.
- 30 terabytes de almacenamiento protegido.
- 400.000 clientes activos con acceso a SQL Server a través del servidor Web IIS o un monitor de transacciones.

SQL Server 2000 es completamente compatible con cualquiera de estos cambios que se realicen en el hardware del servidor para utilizarlos a su favor, para observar esto con más claridad ver la figura siguiente:

Sistema Operativo	Windows 2000 Datacenter	Windows 2000 Advanced Server	Windows 2000 Server	Windows Server 2003 Enterprise Edition	Windows Server 2003 Datacenter Edition	Windows NT 4.0 Server Enterprise	Windows NT 4.0 Server
Número máximo de procesadores compatibles	32	8	4	8	64	8	4
Cantidad máxima de memoria física (RAM) compatible	64 GB	8 GB	4 GB	32 GB (32bit) 64GB (64bit)	32 GB (32bit) 512 GB (64bit)	3 GB	2 GB

Tabla 3.1 Muestra el número máximo de procesadores y memoria física compatibles de SQL Server 2000 Enterprise en diferentes sistemas operativos.

En pruebas comparativas y aplicaciones reales se ha demostrado que una sola CPU puede admitir el acceso de 14 000 usuarios a una BD de 1 terabyte, un nodo de 8 procesadores puede admitir acceso simultaneo de 92 000

usuarios a un SQL Server que administra miles de millones de registros en una matriz de discos de 8 terabytes, y un nodo de 32 CPUs admite el acceso de 290.000 usuarios a una base de datos SQL Server albergada en una matriz de discos de 24 terabytes. El mayor de estos servidores puede procesar más de mil millones de transacciones comerciales al día.

-Escalado horizontal.

La arquitectura de un único nodo puede llegar a tener problemas de cuello de botella que le impide seguir creciendo, derivado de problemas de Entrada/Salida (E/S), muchas veces este problema se puede resolver agregando más memoria física o aumentando la caché para reducir la E/S física. Pero puede darse el caso que los requisitos sobrepasen la capacidad del procesador del hardware de procesamiento simétrico (SMP) de mayor tamaño disponible, este problema se resuelve optando por una arquitectura de escalado horizontal, en la que la carga de trabajo y la BD se dividen entre varios nodos. La arquitectura de escalado puede ser transparente al usuario de la BD y a la aplicación.

SQL Server permite la arquitectura de federación o unión de servidores con vistas distribuidas, donde estos servidores se administran independientemente, pero cooperan en el procesamiento de las peticiones de base de datos de las aplicaciones, cada servidor de la federación se llama servidor miembro.

El primer paso para la confección de una federación de servidores es lograr una buena división de los datos, de esto depende el alto rendimiento de la unión, tratando de que los datos requeridos por una instrucción SQL se encuentren en un solo servidor miembro y de esta forma se disminuye el acceso a servidores remotos. En general se deben dividir las tablas a las que se tiene acceso con más frecuencia, que no tienen que coincidir con la que tienen más datos, por esto es que pueden haber grandes diferencias entre las particiones. Las particiones se realizan generalmente por un campo clave o mediante una función hash aplicada a un campo de la tabla a dividir.

Existen dos formas principales para la división de los datos: partición simétrica y asimétrica.

Partición simétrica.

Este tipo de partición es el modelo ideal, dividiendo las tablas de clave externa junto con la tabla dividida original, de esta forma todos los datos relacionados se encuentran en una misma partición.

Muchas aplicaciones de BD poseen relaciones complejas entre sus datos, por lo que aunque la partición simétrica es la ideal no siempre se puede realizar este tipo de división.

Partición asimétrica.

En este caso solo se pueden dividir algunas tablas de la BD, entonces tenemos dos opciones con las tablas que no se ajustan al esquema de partición, almacenarlas en un mismo servidor o replicarlas en todos los servidores de la unión. El rendimiento de la BD dividida de esta forma es mejor que en el sistema original debido a que la carga de las tablas divididas es distribuida entre los servidores miembros reduciendo la carga en el servidor original.

Existen varias técnicas empleadas a la hora de enfrentarse al problema del diseño de una unión de servidores, algunas de estas son las vistas divididas distribuidas, el Enrutamiento Dependiente de los Datos (DDR, Data Dependent Routing), replicación, creación de particiones mediante un algoritmo Hash, entre otros. Estos métodos se pueden combinar de acuerdo a los objetivos y necesidades de nuestro sistema, en el caso clásico de federación de servidores se utilizan las vistas distribuidas, el DDR y usualmente alguna forma de replicación.

Vistas divididas distribuidas.

Las vistas distribuidas se pueden realizar después que se ha realizado toda la tarea de división de los datos y configuración de los servidores vinculados en cada servidor de la federación.

Una vista dividida combina los datos divididos procedentes de un conjunto de tablas miembro en uno o más servidores, y hace que los datos parezcan proceder

todos de una sola tabla. La vista dividida consiste en una instrucción de selección para cada una de las tablas miembro consolidadas mediante el operador UNION (si se especifica la opción ALL se impide que SQL Server quite las filas duplicadas en el conjunto de resultados). La vista distribuida se define en cada servidor miembro con el mismo nombre. El sistema funciona como si hubiera una copia de la tabla original en cada servidor miembro, aunque cada servidor sólo tiene una parte de la tabla en una tabla miembro y una vista dividida distribuida.

Ejemplo:

Supongamos que tenemos una federación de servidores que cuenta con tres servidores miembros **Server1, Server2 y Server3** y que contienen información sobre clientes de una empresa que están divididos según la región a la que pertenecen. En la vista queremos obtener la información de todos los clientes.

La siguiente vista corresponde a la del servidor miembro Server1.

```
CREATE VIEW Clientes AS
   SELECT * FROM Database.TableOwner.Clientes
UNION ALL
   SELECT * FROM Server2.Database.TableOwner.Clientes
UNION ALL
   SELECT * FROM Server3.Database.TableOwner.Clientes
```

Realizar los mismos pasos en **Server2** y **Server3**.

Las vistas distribuidas ofrecen una gran funcionalidad dado que se pueden consultar como si fueran una tabla más que se encuentra en cualquiera de los servidores de la federación.

Enrutamiento Dependiente de los Datos (DDR).

El objetivo de este método es reducir el tráfico entre los servidores, mediante información para dirigir la consulta al servidor en cuestión, en este método se utiliza código para determinar donde se encuentran los datos necesarios, una vez encontrados se realizan las conexiones pertinentes. En el

DDR la información de cómo encontrar los datos está disponible para la aplicación, una forma sencilla de ubicar esta información es crear una tabla de enrutamiento con la información relativa a la información que contiene cada servidor, la tabla de enrutamiento se puede colocar en uno de los servidores miembros o en otro servidor aparte. No se necesitan servidores vinculados porque las solicitudes se realizan directamente al servidor correspondiente. Por lo general las tablas divididas tienen igual nombre y estructura, pueden poseer diferencias y estas se especifican en la tabla de enrutamiento pero esto solo adiciona complejidad al diseño.

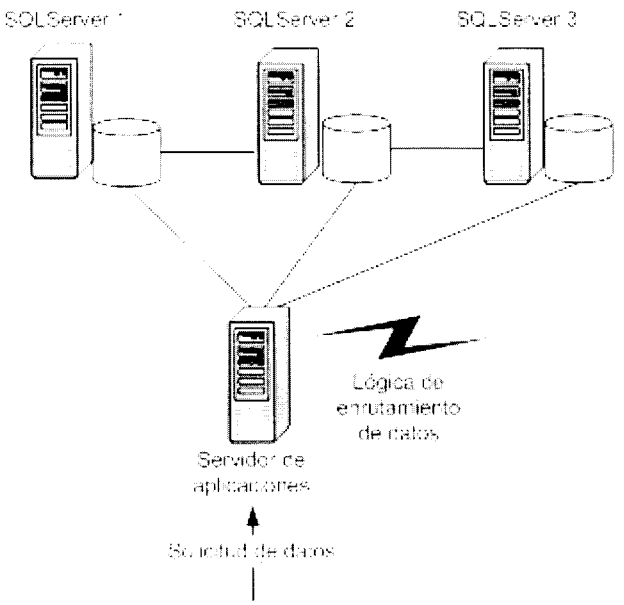


Figura 3.1: Ejemplo de arquitectura para el enrutamiento dependiente de los datos.

Particiones Hash.

Las particiones Hash funcionan de manera similar al DDR, por el hecho de que las solicitudes de enviarán directamente al servidor que contiene los datos. La diferencia es que para la división de los datos se utiliza una función hash que determina en que servidor miembro se guardará el registro. De esta misma forma obtenemos donde se encuentra el registro a la hora de realizarle alguna consulta.

Disponibilidad.

SQL Server ofrece una alta disponibilidad y tolerancia a errores porque proporciona la conmutación por error entre servidores, trasvase de registros y la replicación transaccional.

En un clúster de conmutación por errores los nodos comparten algunos recursos como pueden ser las unidades de discos, y cada nodo puede comunicarse con los demás nodos a través de la red. Los nodos que forman el clúster están periódicamente enviándose mensajes, si se detecta la perdida de uno de ellos se trata al nodo como servidor con error y automáticamente se transfieren los recursos compartidos del clúster y los recursos de aplicación de ese nodo a los otros nodos de la red, los nodos asignados a atender el error producido siguen atendiendo las solicitudes de usuarios del nodo dañado.

SQL Server 2000 implementa el clúster de conmutación por error según las características del servicio Clúster Server de Microsoft (MSCS) de los sistemas operativos Windows NT® 4.0, Windows® 2000 y Windows® 2003.

El trasvase de registros permite mantener una copia de la base de datos en uno o más servidores secundarios, así como promover fácilmente uno de los servidores secundarios para que se convierta en el nuevo servidor principal. Cuando un administrador detecta un error en el sistema realiza un cambio de función convirtiendo un servidor de reserva en servidor principal.

La replicación transaccional usa las tareas del Agente SQL Server para copiar cada modificación realizada en una base de datos de producción a una copia de la base de datos en uno o más servidores secundarios. La replicación transaccional requiere que un administrador detecte un error, deshabilite la replicación y designe uno de los suscriptores como el nuevo servidor principal. Este proceso no es automático y requiere unos minutos para realizarse completamente. Puede tardar bastante más con una base de datos de gran tamaño.

Tras un cambio de función tanto en el trasvase de registro como en la replicación transaccional, los clientes deben conectarse a un servidor distinto con un nombre de servidor y una dirección IP distintos. A diferencia de lo que sucede con el clúster de conmutación por error, los nombres de servidor y las direcciones IP virtuales no se incorporan en estos procedimientos.

SQL Server 2000 Enterprise Edition muestra junto con el sistema operativo Windows 2003 Datacenter Server pruebas satisfactorias comparativas en relación a costo/rendimiento utilizando diferentes arquitecturas.

3.2.2 Oracle.

Oracle es un gestor de base de datos que proporciona la gestión de datos fiable y segura requerida por las aplicaciones críticas OLTP o de data warehousing en las que hay que procesar grandes volúmenes de transacciones online, manteniendo el acceso a la información las 24 horas del día los 7 días de la semana, garantizando los tiempos de respuesta óptimos y la escalabilidad requerida para garantizar que las aplicaciones crezcan al mismo ritmo de los usuarios y la empresa Se ha diseñado por tanto para hacer frente a las exigencias de rendimiento, fiabilidad y escalabilidad que son necesarias para trabajar en la red, tanto para las aplicaciones empresariales tradicionales como para el comercio electrónico en World Wide Web. No sólo aporta una tecnología revolucionaria para mejorar los sistemas en Internet, sino que también convierte Java en el lenguaje de estos sistemas, incluyendo una máquina virtual Java (VM) en el servidor de datos.

Oracle, se ha diseñado con las siguientes características:

- Alta disponibilidad y capacidad de gestión con tablas e índices divididos en particiones.
- Paralelismo mejorado.
- Mayor rendimiento y mejor gestión de aplicaciones de *data warehouse.*
- Procesamiento de transacciones *online* a un nivel comparable al ofrecido por un *mainframe.*

- Mejor administración de la seguridad.
- Mejor soporte de Sistemas Distribuidos.
- Tecnología de objetos y extensibilidad.
- Herramientas de gestión de fácil uso.
- Perfecta migración e interoperatividad entre versiones anteriores.
- Permite Alta accesibilidad brindando la posibilidad de mantener una BD en espera de un error en el acceso de la BD primaria, esta BD se mantiene actualizada de manera automática cada vez que ocurre un cambio en la BD original.

A diferencia de otras soluciones de proveedores, Oracle Database ofrece soporte para muchos estándares del sector a través de las principales arquitecturas de sistemas operativos y de hardware disponibles en la actualidad desde Linux hasta Windows, Unix y OS/390. La portabilidad superior de Oracle Database proporciona mayor capacidad a la organización para cambiar, con facilidad, la infraestructura preferida de hardware y del sistema operativo, asegurando el derecho de una organización de elegir la mejor oferta de precio/desempeño de diferentes proveedores para la actualidad y el futuro. Cualquier organización puede aprovechar al máximo Oracle Database para reducir los costos de implementación iniciales y también para permanecer lo suficientemente flexible para cubrir las necesidades futuras. La elección de Oracle como una solución de base de datos no los compromete con un hardware o sistema operativo particulares. Esto es especialmente útil para los proveedores de software independientes porque pueden realizar el desarrollo en Oracle Database solo una vez y desplegarlo desde cualquier lugar.

3.3-Servidores de BD.

Los servidores son máquinas dedicadas a una tarea específica, que por lo general necesita altas velocidades de procesamiento y/o gran capacidad de almacenamiento. Existen muchos tipos de servidores, por ejemplo: servidores de archivos, de correo, de impresión, etc. Entre estos se encuentran los servidores de

bases de datos, los que con el avance de la tecnología, las comunicaciones y las necesidades actuales de guardar y procesar información, han tenido que evolucionar para soportar las peticiones requeridas.

Los servidores de bases de datos administran el procesamiento de la información entre las aplicaciones y el manejador de bases de datos que se ejecuta en estos.

Las características del servidor dependen del tipo de servicio que este brindará, por ejemplo un servidor de archivos utiliza muy poco el CPU, mientras que un servidor de bases de datos necesita un alto nivel de procesamiento. Entre las principales características que podemos analizar para un servidor de BD tenemos (un servidor de BD puede estar compuesto por más de una máquina, las características mostradas a continuación hacen referencia a un solo equipo):

- Capacidad absoluta de disco duro.
- Arquitectura del bus.
- Versión del BIOS.
- Número de procesadores soportados.
- Capacidad de memoria RAM.
- Capacidad de caché.
- Unidades de disco removibles.
- Tamaños de disco duro soportados.
- Preinstalación y optimización de software.
- Sistemas operativos de red soportados.

3.3.1-Características actuales del servidor de la UCI.

El servidor actual de la UCI está compuesto por un solo nodo que posee las siguientes características técnicas:

- 4 GB de memoria RAM (Random Access Memory, memoria de acceso aleatorio).
- 2 procesadores Xeón a 1 Gh (Gigahertzio, velocidad de procesamiento).
- 5 discos 18 GB (Gigabyte, espacio de almacenamiento).
- El sistema operativo utilizado es Windows 2000 Advanced Server.

3.4 Clusters.

Un clúster es un conjunto de máquinas, conectadas entre sí, generalmente por una red de alta velocidad, que colaboran estrechamente para realizar un trabajo.

El cómputo en Clusters surge como resultado de la convergencia de varias tendencias que incluyen, la disponibilidad de microprocesadores de alto rendimiento más económicos y redes de alta velocidad, el desarrollo de herramientas de software para cómputo distribuido de alto rendimiento, y la creciente necesidad de potencia computacional para aplicaciones en las ciencias computacionales y comerciales.

Los Clusters han evolucionado para apoyar actividades en aplicaciones que van desde súper computo y software de misiones críticas a servidores Web y comercio electrónico, bases de datos de alto rendimiento.

Los clúster en general se pueden agrupar en tres grupos fundamentales:

Clúster de alto rendimiento: es hacer que un número grande de máquinas individuales actúen como una sola máquina muy potente. Este tipo de clúster se aplica mejor en problemas grandes y complejos que requieren una cantidad enorme de potencia computacional. Entre las aplicaciones más comunes de clúster de alto rendimiento se encuentran: el pronóstico numérico del estado del tiempo, astronomía, investigación en criptografía, análisis de imágenes, y más.

Clúster de servidores virtuales: permite que un conjunto de servidores de red compartan la carga de trabajo de tráfico de sus clientes. Al balancear la carga de trabajo de tráfico en un arreglo de servidores, mejora el tiempo de acceso y

confiabilidad. Además, como es un conjunto de servidores el que atiende el trabajo, la falla de uno de ellos no ocasiona una falla catastrófica total.

Clúster de alta disponibilidad: estos tipos de clûster llamados también clúster de redundancia implica tener servidores dentro del clúster que respalden a uno o más servidores de reserva dentro del clúster que actúan de respaldo en caso de la pérdida por algún fallo de los servidores principales.

En su parte central, la tecnología de Clúster consta de dos partes. La primera componente, consta de un sistema operativo confeccionado especialmente para esta tarea, un conjunto de compiladores y aplicaciones especiales, que permiten que los programas que se ejecutan sobre esta plataforma tomen las ventajas de esta tecnología.

La segunda componente es la interconexión de hardware entre las máquinas (nodos) del Clúster. Se han desarrollado interfaces de interconexión especiales muy eficientes, pero comúnmente las interconexiones se realizan mediante una red dedicada de alta velocidad. Es mediante esta interfaz que los nodos del Clúster intercambian entre si asignación de tareas, actualizaciones de estado y datos del programa. Existe otra interfaz de red que conecta al Clúster con el mundo exterior.

3.4.1 Clúster Service de Microsoft.

Este servicio surge de la grande y creciente demanda de sistemas de alta disponibilidad en las organizaciones, ya que las bases de datos y el correo electrónico se hacían esenciales para sus operaciones cotidianas. Como requisitos clave, se identificaron la facilidad de instalación y administración, ya que las pequeñas y medianas empresas suelen tener poco personal de tecnología de la información. Al mismo tiempo, la investigación de Microsoft mostró también una demanda creciente de servidores basados en Windows en grandes organizaciones que exigían un alto rendimiento y disponibilidad.

El Servicio de Clúster Server, que se diseñó originalmente para el sistema operativo Windows NT Server 4.0, se ha adaptado y mejorado para los

sistemas operativos posteriores. El Servicio de Clúster Server permite la conexión de varios servidores en clústeres, con lo que se proporciona alta disponibilidad y fácil administración de los datos y programas que se ejecutan en el clúster. El Servicio de Clúster Server proporciona tres ventajas principales de la tecnología de clúster:

- **Disponibilidad mejorada** al permitir que los servicios y aplicaciones del clúster de servidores continúen en servicio, aunque se produzca un error en un componente de hardware o de software, o durante el tiempo necesario para tareas de mantenimiento.
- **Aumento de la escalabilidad** al admitir servidores que se pueden expandir si se agregan varios procesadores, el número de procesadores varía según el sistema operativo.
- **Administración mejorada** al permitir que los administradores gestionen los dispositivos y recursos de todo el clúster como si estuvieran administrando un único equipo.

Entre los términos fundamentales en la tecnología de clúster se encuentra:

- *Nodo*: cada equipo individual dentro del clúster.
- *Recursos*: son los componentes de hardware y software del clúster que administra el Servicio de Clúster Server. Se dice que un recurso está conectado cuando está disponible y proporcionando servicio al clúster. Entre los recursos se incluyen los dispositivos de hardware físicos, como unidades de disco y tarjetas de red, o los elementos lógicos como direcciones de Protocolo de Internet (IP), aplicaciones completas y bases de datos de aplicaciones.
- *Grupo de recursos*: como su nombre lo indica es un conjunto de recursos administrado por el Servicio de Clúster Server como una única unidad lógica. Contiene todos los elementos que necesita un servidor de aplicaciones y un cliente específicos para poder utilizar la aplicación correctamente. Cuando una operación del Servicio de Clúster Server se

lleva a cabo en un grupo de recursos, la operación afecta a todos los recursos incluidos en el grupo.

El modelo utilizado para diseñar el Servicio de Clúster Server se basó en un modelo *compartir nada* de arquitectura de clúster. Este modelo hace referencia a cómo los servidores de un clúster administran y utilizan los recursos y dispositivos del sistema local y del clúster. En un clúster del modelo compartir nada, cada servidor es propietario de sus dispositivos locales y los administra. En un momento dado, un único servidor es propietario y administra de forma selectiva los dispositivos comunes del clúster.

Servidores virtuales

Las aplicaciones y servicios que se ejecutan en un nodo del clúster de servidores se exponen a los usuarios y estaciones de trabajo como *servidores virtuales*. Ante los usuarios y los clientes, la conexión a una aplicación o servicio que se está ejecutando en un clúster de servidores parece ser un proceso idéntico a la conexión a un único servidor físico. De hecho, la conexión se efectúa con un servidor virtual, que puede estar alojado en cualquier nodo del clúster.

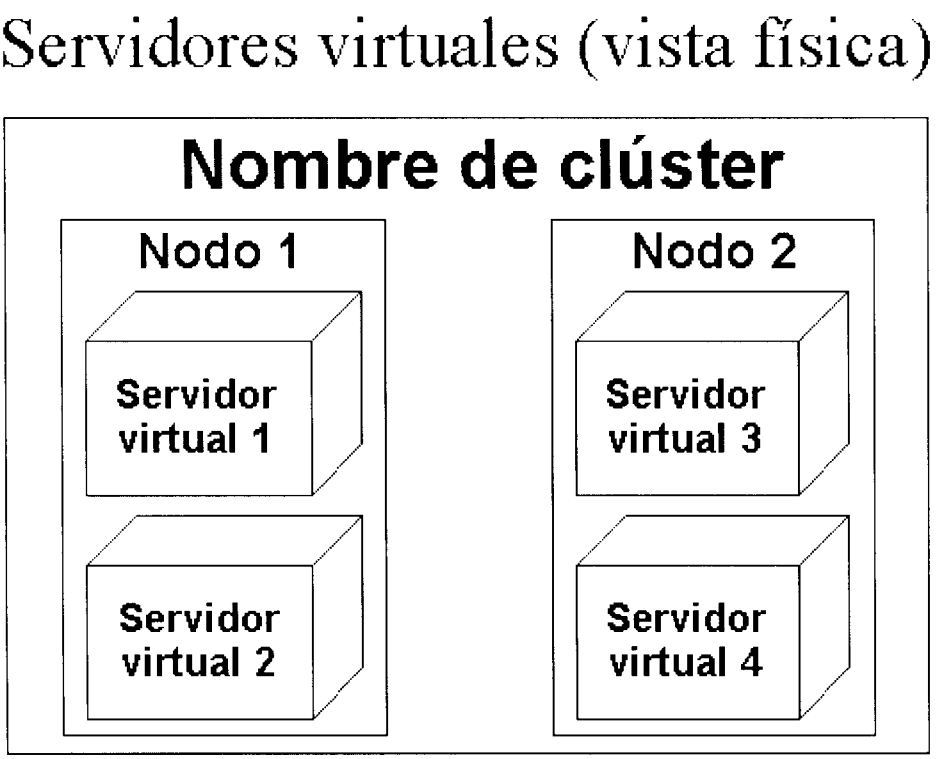


Figura 3.2: Servidores virtuales.

Como se muestra en la figura 3.2 en un clúster pueden existir múltiples servidores virtuales, cada servidor virtual tiene una dirección IP publicada por el Servicio de Clúster, por lo que el cliente se conecta a su aplicación como si fuera un único servidor. En caso de ocurrir un error en uno de los nodos el servicio simplemente asigna el número IP del servidor virtual a otro nodo del clúster que permanezca activo.

El número de nodos por clúster depende del sistema operativo.

	Windows 2000 Advanced Server	Windows 2000 Datacenter Server	Windows 2003 Server
Número de nodos soportados por clúster.	**2**	**4**	**8**

Tabla 3.2 Número de nodos máximos por cluster.

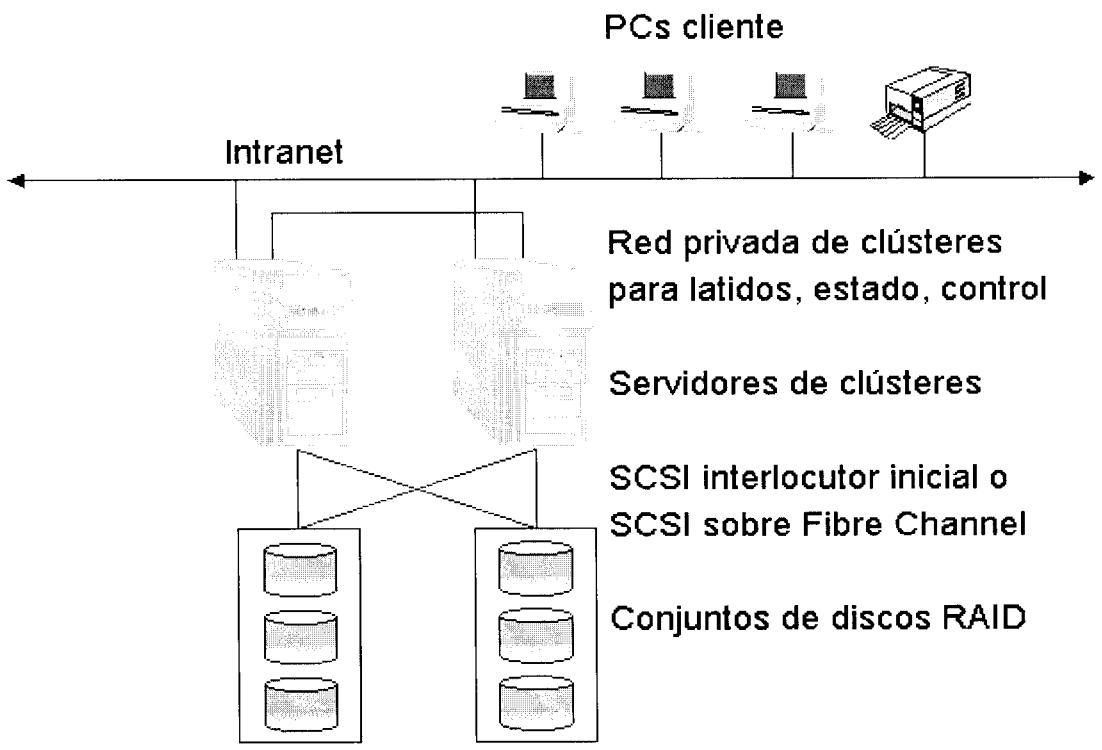


Figura 3.3: Ejemplo de clúster de servidores de dos nodos en los que se ejecuta Windows 2000 Advanced Server.

Sistemas de discos RAID (Conjunto Redundante de Discos Baratos).

Un sistema RAID se compone de un conjunto de discos duros, los cuales son accedidos de forma que el ordenador los ve como si fuese un único disco duro, pero de más capacidad y mayor fiabilidad que cada uno de los discos duros que lo componen por separado. Con este sistema se obtiene mayor capacidad de almacenamiento, redundancia de los datos en caso de fallo de uno de los discos, mayor velocidad de acceso o varios de las anteriores.

Modos en los que puede funcionar un sistema RAID.

- **RAID 0**: en este modo, el sistema RAID combina todos los discos duros en uno solo, cuya capacidad es la suma de todos los discos. De esta forma, sí hacemos un sistema RAID con 4 discos duros de 40GB, el sistema verá un único disco duro de 160GB. Además, los datos son repartidos por todos los discos duros, de forma que la velocidad de acceso también aumenta espectacularmente, pues los datos son leídos de todas las unidades a la vez.
- **RAID 1**: en este modo, todos los discos duros contienen exactamente los mismos datos, de forma que sí uno de ellos falla, se podrá recuperar la totalidad de la información almacenada porque también está copiada en el resto de ellos.
- **RAID 10/0+1:** estos modos son combinación de los dos anteriores; tenemos redundancia mediante un sistema en espejo, y tenemos agrupamiento de discos para conseguir mayor capacidad y velocidad. Ambos ofrecen lo mismo, aunque de una forma ligeramente distinta: mientras que RAID 0+1 es un par redundante de unidades agrupadas, RAID 10 es una agrupación de pares redundantes; la diferencia está en que, en caso de fallo de una unidad, el sistema RAID 0+1 es más vulnerable a sucesivos fallos.
- **RAID 5:** este modo es también una combinación del modo 0 y 1, pero en donde se busca más, el conseguir capacidad extra, que la fiabilidad. Al

igual que en RAID 0, los datos son repartidos entre todas las unidades, pero además, se calcula un bit de paridad. Estos bits de paridad son repartidos a lo largo de cada unidad del array. El mantenimiento de esta tabla baja el rendimiento, pero ofrece un nivel de redundancia que no posee RAID 0. Sí un disco falla en modo RAID 5, es posible recuperar los datos que contenía, gracias a los datos del resto de los discos, la tabla de bits de paridad y un poco de matemática binaria. Los datos de paridad utilizan el espacio de uno de los discos del conjunto.

Un sistema RAID lo podemos tener por hardware o por software, por hardware es mucho más rápido y eficiente, la única desventaja es que es mucho más caro. El RAID por hardware viene implementado en una tarjeta RAID, por lo que el proceso es transparente al SO, mientras que por software, el trabajo lo realiza el propio sistema, consumiendo recursos como memoria y tiempo de CPU.

3.4.2 Clúster de SQL Server 2000 Enterprise Edition.

SQL Server 2000 Enterprise Edition es compatible con los clústeres de conmutación por error, el clúster de conmutación por error reduce el tiempo de inactividad causado por un error del servidor a menos de un minuto, mediante la detección automática de un error del servidor y la iniciación de la conmutación por error a un servidor secundario.

SQL Server permite las siguientes configuraciones de clúster.

- **Clústeres de una sola instancia —** En un clúster de una sola instancia, sólo hay instalado un servidor virtual en el clúster. Los archivos de datos y registro del servidor virtual están instalados en el recurso de almacenamiento compartido para el clúster, y los archivos ejecutables para el servidor virtual están instalados en el recurso de almacenamiento privado para cada nodo. El servidor virtual pertenece al nodo principal, y cada nodo secundario está en estado de espera. Cuando el nodo principal falla o se degrada, se habilita un nodo secundario. Cuando se habilita el nodo, los recursos de SQL Server se inician en el mismo y toman control de los archivos de datos y registro del recurso de almacenamiento compartido. Si

se configura cada nodo con los mismos recursos de hardware y software, el servidor virtual funcionará de forma idéntica en el nodo secundario después de la conmutación por error.

- **Clústeres de varias instancias** — En un clúster de varias instancias, hay instalados dos o más servidores virtuales en el clúster. Los archivos de datos y registro de cada servidor virtual están instalados en un recurso de almacenamiento compartido dedicado a ese servidor virtual. Cuando el nodo principal de un servidor virtual falla o se degrada, el nodo secundario toma control del recurso de almacenamiento compartido del servidor virtual. Puesto que cada servidor virtual tiene un recurso de almacenamiento compartido dedicado, ningún otro servidor virtual se verá afectado por esta conmutación por error.

Para instalar un clúster de varias instancias hay que asegurarse que un nodo pueda realizar el trabajo de los restantes en una conmutación combinada de los restantes, sino, es mejor usar clústeres de una sola instancia cada uno con sus recursos dedicados.

3.4.3 Oracle Real Application Clusters

Oracle Real Application Clusters ofrece una escalabilidad ilimitada y alta disponibilidad con aplicaciones personalizadas o empaquetadas que corran sobre Oracle Database en un entorno de cluster de hardware, manteniendo sin embargo, una manejabilidad y simplicidad de administración como si se tratara de un sistema único.

Oracle Real Application Clusters permite a cualquier aplicación explotar las ventajas de un entorno en cluster: disponibilidad, escalabilidad y rendimiento, sin necesidad de modificar la aplicación. Las aplicaciones pueden tratar Oracle Real Application Cluster como si fuera un entorno simple sobre un sistema único y no necesitan ser modificadas o particionadas para alcanzar la escalabilidad casi lineal de nuestro sistema de cluster de bases de datos. Esto significa que usted puede hacer crecer horizontalmente la capa de base de datos a medida que aumentan las necesidades de hardware de la misma, sin necesidad de modificar la

aplicación. Oracle Real Application Clusters se adapta a la naturaleza cambiante de la carga de base de datos. Puede realizar cambios dinámicos y balanceo de carga a través de los servidores en clúster para alcanzar un rendimiento óptimo.

Oracle Real Application Clusters y Oracle Failsafe ofrecen accesibilidad continua para una base de datos en el caso de una falla del sistema. El failover automático de aplicaciones oculta las fallas de los usuarios para que su trabajo continúe sin interrupciones.

3.4.4 Clústeres de Linux.

Actualmente Linux es una plataforma de computación y súper computación muy consolidada. Es un sistema muy utilizado y mejorado por miles de programadores debido a su facilidad de código libre.

Linux no podía quedarse atrás en el tema de los clústeres de alta disponibilidad, teniendo varias soluciones, en las que algunas destacan por su elegancia y sencillez en comparación con sistemas comerciales.

Algunas soluciones:

Hearbeat: Es una herramienta que permite crear un clúster de alta disponibilidad. De momento el clúster sólo soporta 2 nodos, permite crear grupos de recursos y cambiar estos grupos de recursos fácilmente entre nodos.

Ldirectord y LVS (Linux Virtual Server): LVS permite crear un clúster de balanceo de carga, en el cual hay un nodo que se encarga de gestionar y repartir las conexiones (nodo master LVS) entre todos los nodos slave del clúster. El servicio de datos debe residir en todos los nodos slave. LVS puede llegar a soportar sin problemas hasta 200 nodos slave. Ldirectord es un demonio que se ejecuta en el master LVS, que se encarga de testear el servicio de datos de los nodos slave y eliminarlos e insertarlos en el clúster dinámicamente, si surge algún problema o sí se repone el servicio según sea el caso.

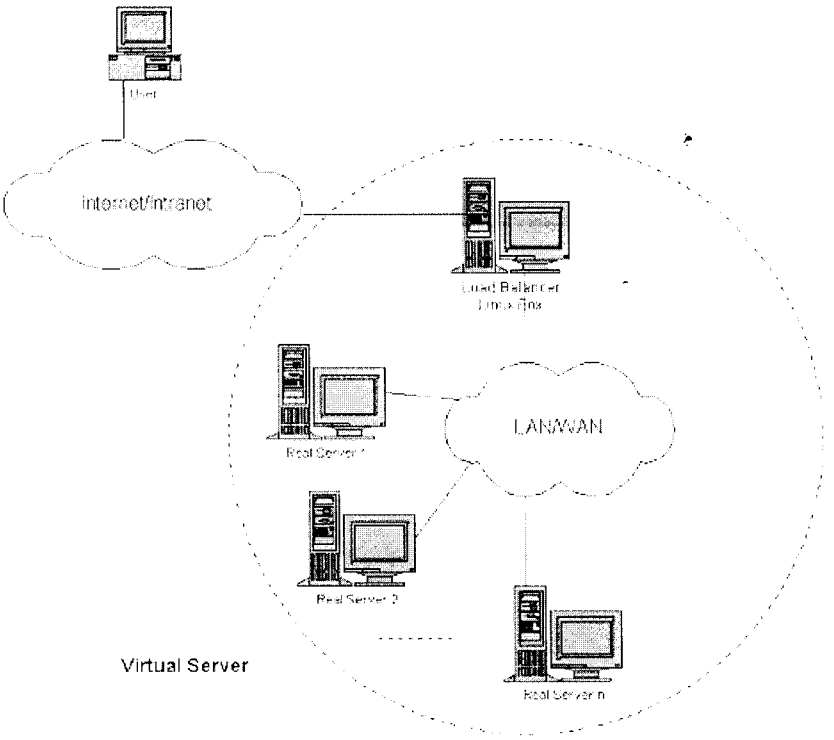


Figura 3.4: Arquitectura general de clúster de balance de carga utilizando LVS.

LVS permite diferentes configuraciones:

- ***NAT* (*Network Address Translation*)**: El nodo master que es el encargado del balanceo de carga recibe la petición del cliente, el paquete es reescrito y enviado a uno de los servidores, el servidor procesa la petición y devuelve los resultados al master, este reescribe los paquetes de respuesta y los envía al cliente.
- ***IP Tunneling***: El Load Balancer recibe la petición del cliente, el paquete es encapsulado (poner un datagrama IP dentro de otro) y reenviado a uno de los servidores, el servidor desencapsula el paquete, procesa la petición y envía la respuesta directamente al cliente.
- ***Direct Routing:*** El nodo master recibe la petición del cliente, se elige el servidor adecuado y se enruta el paquete hacía él mediante su dirección MAC (*Media Access Control, dirección física que se encuentra en la tarjeta de red, cada tarjeta tiene un identificador único*), el servidor procesa la petición y devuelve los datos al cliente directamente.

UltraMonkey: Es una solución creada por VA Linux que se basa en LVS y Heartbeat para ofrecer clústeres de alta disponibilidad y balanceo de carga (ver Figura 3.5). El nodo master LVS se pone en alta disponibilidad ya que es el único punto de ruptura. Además, incorpora una interfaz para configurar el clúster.

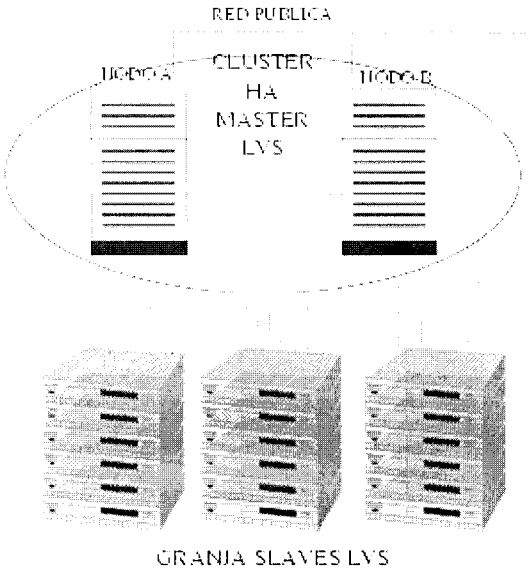


Figura 3.5: Arquitectura de clúster utilizando UltraMonkey.

Existen muchas otras soluciones como son: Pirahna, Kimberlite, Mosix, etc.

3.5-Propuesta de Infraestructura.

Al analizar una propuesta de infraestructura para el futuro Centro de Datos de la UCI, lo primero que debemos realizar es un breve estudio de las diferentes BD que se desean albergar en el mismo y tener una aproximación de los posibles usuarios de las mismas.

En la actualidad las BD del centro están ubicadas en un único nodo que ejecuta Windows 2000 Advanced Server y SQL Server 2000 Enterprise Edition. Estas BD guardan datos de aplicaciones que se utilizan solamente dentro del centro. La UCI, dentro de su proceso de informatización, irá desarrollando numerosas aplicaciones, las cuales pueden hacer uso de una BD ya construida o nueva. Por esta razón el número de BD puede crecer a medida que el centro avanza por el largo camino de automatizar la mayor cantidad de servicios posibles.

Está previsto que el centro llegué a tener, como máximo, alrededor de ***15000*** personas y el número de máquinas debe superar las ***6000***. Un dato importante a analizar es que las BD solo serán accedidas por usuarios del centro, por lo que se puede llegar a aproximar el número de conexiones máximo con el número de máquinas, esto es una gran ventaja porque conociendo el número máximo de conexiones es posible diseñar un sistema completamente disponible y de alto rendimiento.

Después de realizar un detallado análisis a la variedad de tecnologías disponibles para el montaje de un centro de datos que potencie una alta disponibilidad y escalabilidad, la mejor opción la muestra la creación de un clúster de servidores de alta disponibilidad. En este tipo de sistema se posee por lo general recursos redundantes para la mayoría de puntos críticos dentro del sistema, logrando de esta manera que la pérdida por algún fallo de uno de ellos no produzca un paro en el sistema. La tecnología de clúster está siendo utilizada por la mayoría de las empresas que brindan servicios 24x7x365, o sea, que tienen que estar disponibles todo el año.

Para dar una infraestructura más detallada en está primera investigación podemos basarnos sobre Microsoft Clúster Service debido a que el sistema operativo Windows es el más difundido en la UCI y la mayoría de las aplicaciones están basadas en esta plataforma, además de poseer actualmente las BD sobre el SGBD SQL Server 2000.

Anteriormente, cuando realizábamos un estudio sobre SQL Server 2000, mencionamos que en una prueba comparativa con un nodo de 8 procesadores el sistema pudo atender hasta 92 000 usuarios conectados simultáneamente, en una BD que administra miles de millones de registros en una matriz de discos de 8 terabytes y que una sola CPU puede admitir el acceso de 14 000 usuarios a una BD de 1 terabyte. Esto nos da una medida de la potencia del sistema y de lo que podemos diseñar sin llegar a realizar un gasto excesivo de recursos.

Analizando todo lo mostrado hasta el momento queda demostrado que una arquitectura compuesta de un clúster de 4 nodos que ejecutan Windows Server

2003 por conmutación por error permite de manera eficiente responder a las necesidades de escalabilidad y disponibilidad de la UCI.

1) ***La escalabilidad*** se logra fácilmente por la adición de más procesadores y memoria interna a las máquinas existentes en el clúster o actualizando a un nodo más grande (escalabilidad vertical) y/o adicionando máquinas al clúster. El sistema operativo Windows Server 2003 Enterprise Edition hace uso eficiente de hasta 8 procesadores y 32 GB de memoria interna para la versión de 32 bits y 64 GB para la de 64 bits, y la versión Datacenter de hasta 64 procesadores y 32 y 512 GB para versiones de 32 y 64 bits respectivamente. Logrando con esto rendimientos altos por cada máquina por separado. En el caso de la adición de máquinas utilizando este sistema operativo el clúster puede tener hasta ocho nodos.

2) ***La disponibilidad*** del sistema se logra por la configuración de conmutación por error del clúster, en caso de error en un nodo por software o hardware el clúster automáticamente distribuye las aplicaciones en un nodo activo restante. En cada nodo puede existir más de un servidor virtual, cada uno de estos en caso de error posee su política de preferencia sobre que nodo reiniciarse en caso de error. Esto brinda una gran flexibilidad al sistema, no es obligatorio que todos los nodos tengan hardware idéntico, y existen aplicaciones que requieren una mayor carga, siendo un requisito indispensable, mantener el rendimiento en caso de error para que el usuario no sienta un cambio considerable.

3) ***La disponibilidad de los datos*** se logra con el sistema de discos RAID que puede ser configurado en modo 5 ó 10 para obtener una mayor capacidad de recuperación ante fallos. En general, ambos modos brindan los mismos beneficios de redundancia de los datos, pero el modo 10 lo hace a costa de necesitar mayor espacio en disco y el 5 hace mayor uso de la CPU, lo que implica menor velocidad en el sistema, principalmente sí se utiliza la opción de RAID por software. Además, de esta opción se puede utilizar alguna de las opciones brindadas por SQL Server 2000 como el trasvase de registros para un

servidor secundario, teniendo de esta forma una copia completa de las BD en algún almacenamiento externo al cluster.

Para el montaje inicial podemos contar preferiblemente con nodos que cuenten con al menos dos procesadores con una velocidad superior a 1 GH y memoria interna superior a los 5 GB. Estos nodos deben poseer un hardware extensible, que permita adicionar más procesadores y memoria interna en un futuro brindando la posibilidad de escalado vertical sin tener que comprar un nodo más grande.

Entre otros detalles podemos mencionar la redundancia de fuentes de electricidad y conexiones de red, evitando la salida del sistema por un fallo eléctrico o por caída de la red principal.

Para encontrar el hardware específico que necesitamos, sí el clúster se implementa con software de Microsoft, como proponemos en este caso, podemos consultar el sitio http://www.microsoft.com/whdc/hcl/default.mspx, el cual muestra la lista de hardware compatibles, los servidores mostrados en el sitio por cuestiones de mercado son productos que han pasado duras pruebas de escalabilidad y disponibilidad.

En el sitio se encuentra una arquitectura con características similares a la solución propuesta:

- Sistema Operativo Windows Server 2003, Enterprise Edition.
- Número de nodos: 4.
- Almacenamiento externo RAID.
- Características de cada nodo: Marca **IBM eServer xSeries 335-8676 (Dual 3.06Ghz),** modelo CPU INTEL XEON 512KB L2 HYPER THREADING.

En el clúster se pueden instalar el número de servidores virtuales deseados y varias instancias de SQL Server, sí es posible, es recomendable poseer, dentro del clúster, un nodo que permanezca como nodo pasivo que sólo entrará en trabajo sí ocurre alguna conmutación de los nodos restantes. En el peor de los casos, el trabajo puede recaer sobre un solo nodo, por lo que algunas veces, en

vez de aumentar el número de nodos de un clúster es recomendable crear uno nuevo, poseyendo cada uno sus propios recursos.

En un futuro se puede realizar un cambio en la infraestructura del Clúster, basándose en el SGBD Oracle el cual es compatible con varias plataformas y muestras grandes potencialidades en este campo junto con el sistema operativo Linux.

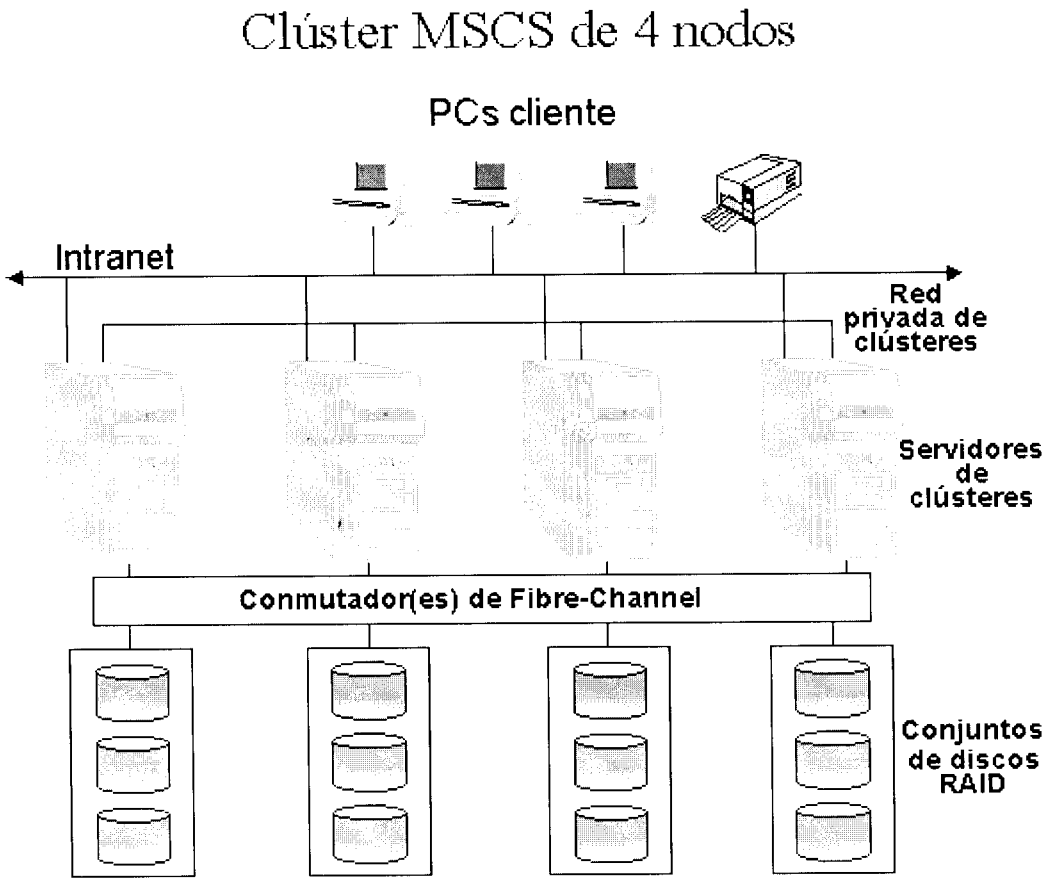


Figura 3.6: Ejemplo del Clúster.

3.6-Conclusiones.

Los cluster de computadoras se han venido imponiendo en los últimos tiempos, brindando una solución económica y fiable a los problemas de alta disponibilidad. Esto ha sido consecuencia de los altos precios adquiridos por los grandes servidores, y por el avance de la tecnología en las redes de alta velocidad y microprocesadores de alto rendimiento más económicos. Siendo una forma más barata y que brinda las mismas posibilidades a pequeñas y medianas empresas que no tienen acceso a las grandes máquinas.

Capítulo 4

Herramientas OLAP

4.1 Introducción a los Almacenes de Datos (Data Warehouse) y OLAP.

El almacenamiento de datos (*data warehousing*) y el procesamiento analítico en línea (*online analytical processing)* son elementos esenciales en el soporte de decisiones, que se están convirtiendo de forma creciente en un foco de la industria de las bases de datos.

Los Almacenes de Datos (Data Warehouse, AD) surgen con la necesidad de las empresas de disponer de bases de datos que permitan extraer conocimientos de la información histórica de la organización, con el objetivo de disponer de Sistemas de Información de apoyo a la toma de decisiones (realizar análisis de la empresa, previsiones de evolución, diseño de estrategias, etc.).

Las principales características de un AD son:

- Mientras que las BD operacionales (OLTP, *Online Transaction Processing*) son orientadas al proceso, los AD están orientados a la información relevante de la organización, debido a que se diseña para consultar de una manera eficiente, la información relativa a las actividades principales de la empresa, no para soportar los procesos que se realizan.

- La información es integrada, integra datos recogidos de diferentes sistemas operacionales de la organización y/o fuentes externas.

- Variable en el tiempo, debido a que los datos son relativos a un período de tiempo y son incrementados periódicamente.

- La información es no volátil porque los datos no son actualizados, sólo son incrementados.

Almacena datos actuales	Almacena datos históricos
Almacena datos de detalle	Almacena datos de detalle y datos agregados a distintos niveles
Bases de datos medianas (100Mb-1Gb)	Bases de datos grandes (100Gb-1Tb)
Los procesos (transacciones) son repetitivos	Los procesos no son previsibles
El número de transacciones es elevado	El número de transacciones es bajo o medio
Tiempo de respuesta pequeño (segundos)	Tiempo de respuesta variable (segundos-horas)
Dedicado al procesamiento de transacciones	Dedicado al análisis de datos
Orientado a los procesos de la organización	Orientado a la información relevante
Soporta decisiones diarias	Soporta decisiones estratégicas
Los datos son dinámicos (actualizables)	Los datos son estáticos
Sirve a muchos usuarios	Sirve a técnicos de dirección

Tabla 4.1 Diferencias entre un sistema operacional (OLTP) y un AD.

La tecnología OLAP permite un uso más eficaz de los almacenes de datos para el análisis en línea, lo que proporciona respuestas rápidas a consultas analíticas complejas e interactivas. Los modelos de datos multidimensionales de OLAP y las técnicas de agregados de datos organizan y resumen grandes cantidades de datos para que puedan ser evaluados con rapidez mediante el análisis en línea y las herramientas gráficas. La respuesta a una consulta realizada sobre datos históricos a menudo suele conducir a consultas posteriores en las que el analista busca respuestas más concretas o explora posibilidades. Los sistemas OLAP proporcionan la velocidad y la flexibilidad necesarias para dar apoyo al analista en tiempo real.

4.2 ¿Que son las herramientas OLAP?

El término OLAP (*Online Analytical Processing*) sugiere un modelo de aplicaciones orientadas fundamentalmente a las consultas complejas que involucran a una gran cantidad de datos, de forma que bajo su punto de vista funcional son herramientas de servicio de datos bajo una idea de optimizar la consulta y con abstracción del aspecto transaccional. Es la tecnología que permite acceder de una manera eficiente a la información de los AD, brindándole numerosas ventajas al usuario, por ejemplo:

- Presenta al usuario una visión multidimensional de los datos (esquema multidimensional), facilitando la selección, recorrido y exploración de los datos.
- El usuario formula consultas a la herramienta OLAP seleccionando atributos de este esquema multidimensional sin conocer la estructura interna (esquema físico) del almacén de datos.
- Brinda un lenguaje analítico de consulta que proporciona la capacidad de explorar las complejas relaciones existentes entre los datos empresariales.
- Precálculo de los datos consultados con más frecuencia.

4.3 Bases de Datos Multidimensionales.

Como vimos anteriormente las herramientas OLAP presentan al usuario los datos de una manera multidimensional para facilitar el trabajo y análisis de estos, pero, ¿Qué es una Base de Datos Multidimensional (BDM)?

Una base de datos multidimensional representa una abstracción acerca de las diferentes visiones de un conjunto de datos y las relaciones que se pueden establecer sobre los mismos. Ofrecen, por tanto, una visión orientada al análisis de las relaciones entre categorías de datos. Las Bases de datos multidimensionales proveen la consolidación y cálculos, según las diferentes vistas posibles, es decir, según las diferentes dimensiones que se pueden

configurar de acuerdo a la estructura definida, de forma que el usuario puede pivotar según estas dimensiones. Cada celda intersección de planos representa un valor, normalmente de frecuencia de la relación entre valores de las categorías que representan la intersección.

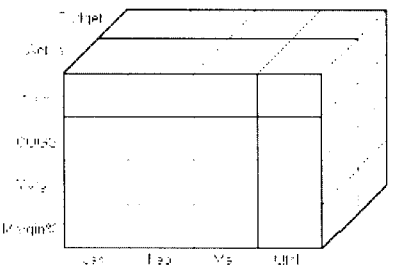


Figura 4.1: BD tridimensional. Forma de cubo.

4.4 Cubos Multidimensionales.

Los cubos son los principales objetos del proceso analítico en línea (OLAP), una tecnología que proporciona rápido acceso a los datos de un almacén de datos. Un cubo es un conjunto de datos que normalmente se construye a partir de un subconjunto de un almacén de datos y se organiza y resume en una estructura multidimensional, definida por un conjunto de ***dimensiones*** y ***medidas***. La estructura del cubo está determinada por sus dimensiones, mientras que las medidas brindan valores numéricos importantes para el análisis de la actividad tratada. Las posiciones de las celdas en el cubo se definen mediante la intersección de los miembros de la dimensión, y los valores de las medidas se agregan para proporcionar los valores de las celdas.

Las dimensiones representan el más alto nivel de tratamiento de las diferentes categorías de datos a analizar, siendo la visión de cada uno de los ejes coordenados del espacio multidimensional. Están compuestas por miembros (atributos de dimensión) que la caracterizan, que pueden estar organizados de acuerdo a una jerarquía de niveles, facilitando posteriormente la navegación por los diferentes niveles según el grado de simplicidad que se desea en la consulta. En cada nivel puede haber uno o más miembros de la dimensión. Las

dimensiones caracterizan cada una de las vistas de la actividad analizada en el cubo y brindan los diferentes datos a combinar.

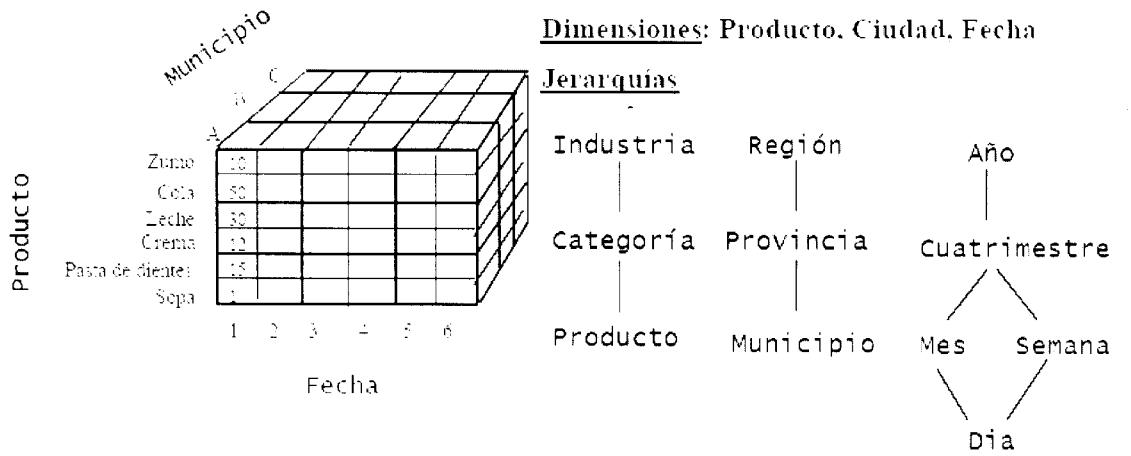


Figura 4.2: Ejemplo de cubo, con sus dimensiones y jerarquías.

Como se muestra en la figura 4.2 cada dimensión tiene su jerarquía de acuerdo al nivel de detalles hasta el cual desea el usuario analizar los datos. La dimensión ***Tiempo*** es de mucha utilidad debido a que la mayoría de las consultas para la toma de decisiones dependen de un intervalo de tiempo.

Las celdas son la estructura mínima de almacenamiento formada por la intersección de un valor de cada una de las dimensiones que componen el hipercubo. Puede contener o no contener datos.

4.4.1 Estructura del Cubo.

La estructura de un cubo se define por medio de sus medidas y dimensiones. Derivan de tablas del origen de datos del cubo. El conjunto de tablas del que derivan las dimensiones y medidas de un cubo se denomina esquema del cubo. Cada esquema de cubo consta de una sola tabla de hechos y de una o más tablas de dimensiones. Las medidas del cubo derivan de columnas de la tabla de hechos. Las columnas de las dimensiones del cubo derivan de columnas de las tablas de dimensiones.

- **Tabla de hechos:** Está compuesta por las claves extranjeras de cada una de las tablas de dimensiones y por las medidas. Las medidas son el

objetivo del cubo, siendo los datos más importantes requeridos por el usuario a la hora del análisis, conteniendo la información de la actividad referenciada por el cubo. Los datos de la tabla de hechos son por lo general numéricos, permitiendo cálculos estadísticos y agregaciones.

- **Tablas de dimensión:** Describen los atributos de los objetos involucrados en la actividad analizada. Por lo general está compuesta por un identificador del objeto y sus propiedades.

Para el diseño de la estructura del cubo existen dos modelos fundamentales: esquema en estrella (*star schema*) y el esquema copo de nieve (*snowflake schema*).

Esquema en estrella.

En este tipo de estructura cada tabla de dimensión solo se relaciona con la tabla de hechos. Por lo general la tabla de dimensión se encuentra desnormalizada.

El esquema en estrella tiene como ventajas que proporciona una mayor rapidez a las consultas dado que realiza menos uniones para poder dar un resultado, es fácil de usar y entender y se puede extender mejor a cambios futuros. La desventaja es que posee redundancia de los datos por lo que ocupa mayor espacio en disco.

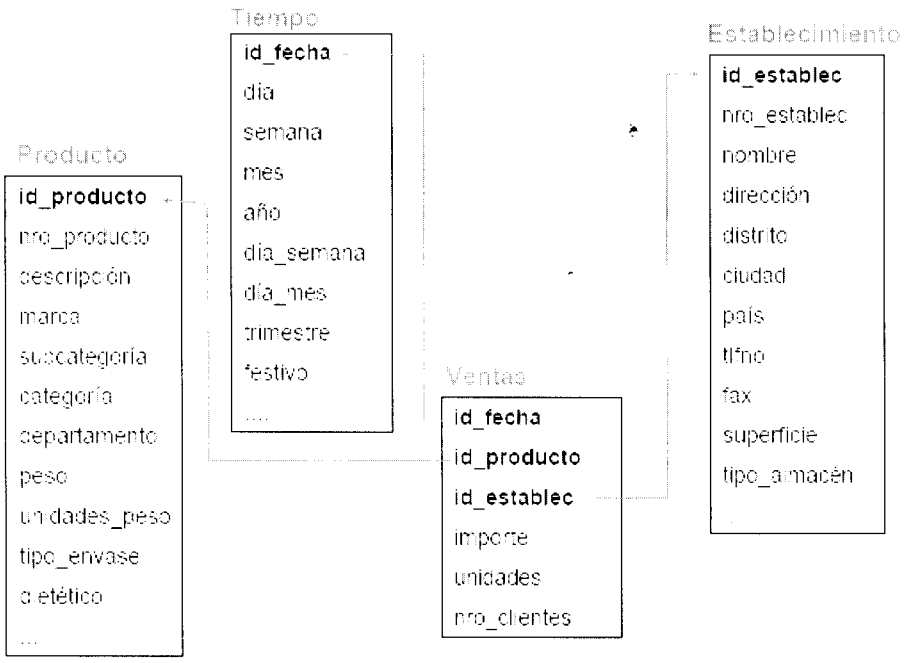


Figura 4.3: Ejemplo de esquema en estrella.

Esquema en copo de nieve.

En este tipo de estructura a diferencia del esquema en estrella las tablas de dimensiones pueden estar relacionadas con otras tablas de dimensión, permitiendo la normalización de las tablas de dimensiones.

De esta forma se elimina la redundancia de los datos, pero el tiempo de respuesta a las consultas es mayor comparado con el esquema en estrella.

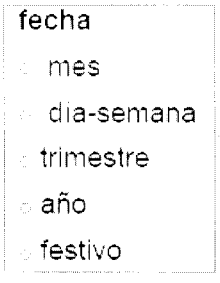


Figura 4.4: Dimensión Tiempo desnormalizada.

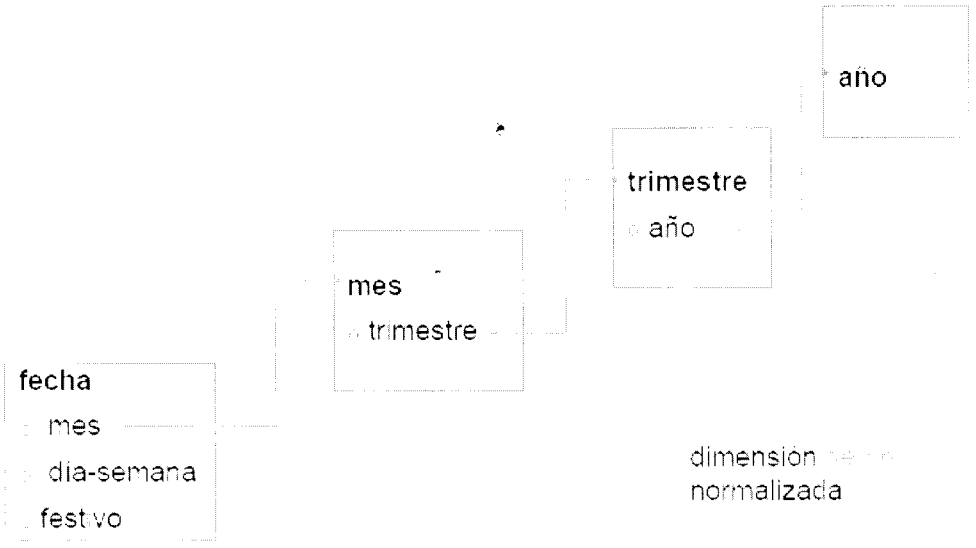


Figura 4.5: Dimensión Tiempo normalizada.

4.4.2 Datos agregados.

Los agregados son resúmenes de datos precalculados que mejoran el tiempo de respuesta a las consultas por el simple hecho de tener preparadas las respuestas antes de que se planteen las preguntas. Por ejemplo, la respuesta a una consulta que solicita el total de ventas semanales de una determinada línea de productos y que se realiza en una tabla de hechos de un almacén de datos que contiene cientos de miles de filas de información, puede llevar mucho tiempo sí hay que explorar la tabla de hechos para calcular la respuesta. Por el contrario, la respuesta podría ser casi inmediata sí los datos de resumen para la respuesta a esta consulta se han calculado previamente. El cálculo previo de los datos de resumen es la clave para obtener respuestas rápidas en la tecnología OLAP. Sí se calculan previamente todos los posibles agregados a un cubo, se obtiene el tiempo de respuesta más corto posible para todas las consultas. Sin embargo, el tiempo de almacenamiento y el tiempo de proceso necesarios para todos los agregados puede ser sustancial.

El total de agregados de un cubo no solo depende de las medidas y las dimensiones, sino también, del número de niveles en las jerarquías de las dimensiones y de la cantidad de miembros en cada nivel.

Los datos agregados son una de las principales ventajas que presenta la tecnología OLAP sobre los sistemas relacionales, proporcionando una respuesta inmediata a un resumen que puede incluir miles de filas. Los datos agregados pueden ser cualquier tipo de fórmula que generalmente involucra a varias medidas del cubo.

4.4.3 Operadores de refinamiento de consultas.

Utilizando la tecnología OLAP las vistas son fácilmente configurables por el usuario, las dimensiones no se encuentran estáticamente dispuestas en filas y columnas, el usuario puede modificar la forma y el nivel de agregación de los datos apelando a los operadores de refinamiento OLAP, como son: *drill-down*, *drill-up*, *slice*, *dice, pivot* y *nesting*.

Como ya sabemos, los miembros de una dimensión pueden estar organizados jerárquicamente en niveles, el operador *drill-down* permite llevar una consulta a un nivel más detallado, rastreando por los diferentes niveles desde un nivel más general a uno más minucioso. Por ejemplo sí tenemos una dimensión Región con la siguiente jerarquía Región→País→Estado, y tenemos la consulta Ventas por región, utilizando el operador *drill-down* podemos obtener las Ventas por Países. *Drill-Up* realiza la función opuesta del operador anterior, consolidando y agregando los datos obtenidos para dar un reporte más general.

Las operaciones *Slice y dice* permiten seleccionar y proyectar datos de un corte bidimensional de un cubo.

El operador *pivot* permite reorientar las dimensiones en un informe o consulta, permitiendo cambiar una dimensión que se encuentra en un informe como columna que aparezca como fila y viceversa o girar, cambiar las filas por las columnas.

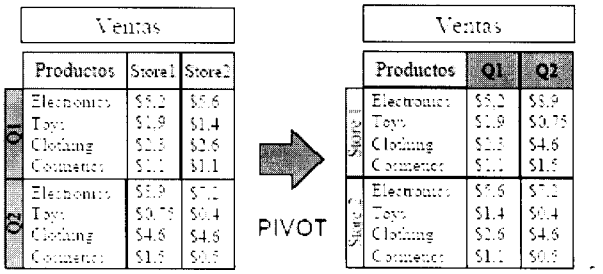


Figura 4.6: Aplicación del operador *pivot.*

El operador *nesting* permite anidar una dimensión dentro de otra, desplegando una dimensión para cada uno de los miembros de otra dimensión.

4.4.4 Procesar cubos.

Esta operación constituye la parte final del engranaje para completar los datos requeridos para nuestro análisis. Con el procesamiento de un cubo se leen las tablas de dimensiones para llenar con datos actuales cada uno de sus niveles, se lee la tabla de hechos, se calculan los agregados y se almacena todo el resultado en el cubo. Terminado este proceso el cubo OLAP está listo para ser consultado.

El procesamiento puede ser de tres formas distintas:

- Procesamiento completo.
- Actualización incremental.
- Actualizar los datos.

Procesar un cubo

Procesar es el término que se utiliza para una carga completa del cubo. Se leen todos los datos de las dimensiones y de la tabla de hechos, y se calculan los agregados especificados. Se debe procesar un cubo cuando su estructura sea nueva o cuando se hayan modificado sus dimensiones o medidas. El proceso de un cubo puede llevar mucho tiempo sí existe una tabla de hechos de gran tamaño y hay muchas dimensiones con gran cantidad de niveles y muchos elementos en cada nivel.

Siempre que exista un cambio en la estructura del AD que afecten al cubo o se cambie la estructura del cubo, se debe procesar nuevamente el mismo. Existen otros cambios como modificaciones en los datos agregados o inserciones de nuevos datos que no requieren un proceso completo, estos cambios se pueden actualizar al cubo existente mediante las opciones de procesamiento *Actualización incremental* o *Actualizar los datos*.

Actualización incremental

Una actualización incremental es adecuada cuando se van a agregar nuevos datos al cubo, pero los datos existentes no cambian y la estructura del cubo sigue siendo la misma. La opción *Actualización incremental* agrega nuevos datos y actualiza agregados; no requiere un procesamiento completo del cubo.

Una actualización incremental no afecta a los datos existentes que ya se han procesado. Requiere bastante menos tiempo que la ejecución de un procesamiento. Una actualización incremental se puede llevar a cabo mientras los usuarios consultan el cubo, una vez finalizada la actualización, los usuarios tienen acceso a los datos adicionales sin necesidad de desconectarse y volverse a conectar.

Actualizar los datos de un cubo

La opción *Actualizar los datos* hace que se borren y vuelvan a cargar los datos de un cubo, y se vuelvan a calcular los agregados. Esta opción es adecuada cuando se han modificado los datos subyacentes contenidos en el almacén de datos, pero se ha conservado la estructura del cubo. Esta opción es más rápida que procesar el cubo porque no es necesario volver a diseñar las tablas de agregados y la estructura del cubo.

4.5 Tipos de almacenamiento OLAP.

Existen tres formas básicas para el almacenamiento de los cubos multidimensionales: ROLAP (Relacional OLAP), MOLAP (Multidimensional OLAP) y HOLAP (OLAP Híbrido).

- ROLAP: Toda la información del cubo es almacenada en una BD relacional.

- MOLAP: los datos fuentes del cubo son almacenados junto con sus agregaciones en una estructura multidimensional de alto rendimiento, los datos guardados de esta forma ofrecen excelentes rendimientos.

- HOLAP: Combina los dos tipos anteriores, la agregación de datos es almacenada en una estructura multidimensional y los datos fuentes del cubo en una BD relacional.

El sistema de almacenamiento MOLAP proporciona los tiempos de respuesta a consultas más rápidos, que dependen únicamente del porcentaje y del diseño de los agregados del cubo. En general, MOLAP es más apropiado para cubos de uso frecuente y que necesitan tiempos de respuesta muy cortos. MOLAP brinda espectaculares tiempos de respuesta a costa de mayor necesidad de espacio y retardo en las modificaciones.

A diferencia del almacenamiento MOLAP, ROLAP no almacena una copia de los datos base, sino que tiene acceso a la tabla de hechos originales cuando es necesario para responder a consultas. Las respuestas a consultas ROLAP suelen ser más lentas que aquellas que se realizan con las otras dos estrategias de almacenamiento. Un uso típico de ROLAP es el acceso a grandes conjuntos de datos consultados con poca frecuencia, tales como datos históricos de años no recientes.

HOLAP es el equivalente de MOLAP, para las consultas que tienen acceso a los datos de resumen. Las consultas que tienen acceso a datos base, por ejemplo una consulta que aumenta el nivel de detalle hasta un hecho simple, deben recuperar los datos de la base de datos relacional y no se ejecutarán con tanta rapidez como cuando los datos bases están almacenados en la estructura MOLAP. Los cubos almacenados como HOLAP tienen un tamaño menor que los cubos MOLAP equivalentes y responden con mayor rapidez que los cubos ROLAP a consultas relativas a datos de resumen. El almacenamiento HOLAP suele ser

adecuado para cubos que requieren tiempos cortos de respuesta para consultas realizadas en resúmenes basados en grandes cantidades de datos base.

En el mercado existen numerosas herramientas para el soporte de la tecnología OLAP, como por ejemplo OlapX Software, MicroStrategy 7, Cognos PowerPlay® 7.3, entre otras. Ahora veremos algunas características sobre Analysis Services que viene incluida con SQL Server 2000 Enterprise Edition.

4.6 Ejemplo de Herramienta OLAP. Analysis Services (AS) de SQL Server 2000.

Analysis Services es una nueva parte incluida en SQL 2000 que amplia el antiguo Servicio OLAP que se introdujo en la versión 7.0 de SQL Server. Este servicio contiene todas las características que hasta el momento hemos visto poseen los cubos multidimensionales, pero a su vez, incorpora numerosas funcionalidades y opciones que aumentan el poder de análisis de los cubos, ayudando en consecuencia a los analistas en su trabajo.

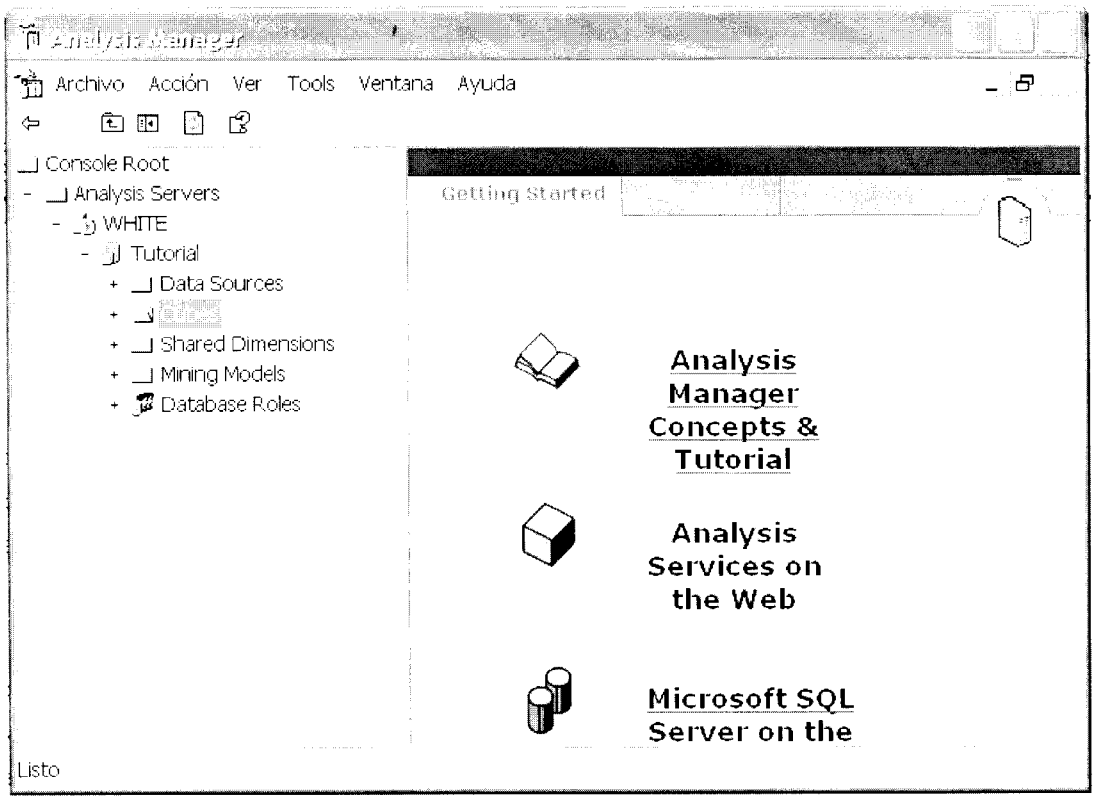


Figura 4.7: Analysis Manager, vista principal.

Microsoft con esta herramienta ofrece la posibilidad de crear, gestionar y consultar cubos de datos de manera inteligente. El OLAP Server soporta MOLAP, ROLAP y HOLAP; niveles variables de agregación para optimizar el rendimiento de las consultas frente al espacio de almacenamiento; datos origen en esquemas dimensionales o relacionales; particionado de cubos para posibilitar consultas contra orígenes de datos distribuidos y heterogéneos; análisis de utilización que permiten examinar las consultas con problemas y reconstruir las agregaciones para ajustar dichas consultas; posibilidad de deshacer para el desarrollo de análisis de hipótesis; posibilidad de actualización incremental; y una interfaz de OLE DB ampliada para OLAP.

Analysis Services permite crear particiones remotas de un cubo de esta forma distribuyendo la utilización de la CPU y la memoria porque al procesar una partición remota la mayor parte del trabajo lo realiza el Analysis Services remoto. Las particiones se pueden almacenar mediante un modo distinto (MOLAP, ROLAP, HOLAP). Las particiones de un cubo son invisibles para el usuario, sin embargo, es importante que las particiones se definan de tal manera que contengan datos mutuamente exclusivos. Un cubo puede proporcionar resultados incorrectos a algunas consultas, sí una parte de los datos del cubo está incluida en más de una partición. Las particiones de un cubo se pueden almacenar en varios servidores para proporcionar un método de almacenamiento en cubos basado en clústeres. Dos particiones de un cubo pueden mezclarse en una única partición que, a su vez, puede combinarse con otra partición y así sucesivamente hasta que quede una única partición. Por ejemplo, se puede mezclar cuatro particiones, cada una de las cuales contienen datos correspondientes a un trimestre, en una única partición que contenga los datos de todo el año.

Esta herramienta permite los cubos en tiempo real, normalmente cuando cambian los datos subyacentes de un cubo se debe procesar nuevamente para actualizarlo, lo cual provoca bloqueos del cubo durante períodos de tiempo en los que los usuarios no pueden acceder, esta opción resuelve este problema y es de gran utilidad en los cubos basados en tablas de hechos o dimensiones que cambian frecuentemente, habilitando las particiones y dimensiones OLAP para

actualizarse automáticamente cuando ocurre algún cambio en los datos origen del cubo.

La opción cubos vinculados permite crear un cubo en un Analysis Services y definirse como vinculado en otros Analysis Services, de esta forma los usuarios se pueden conectar a cualquiera de los AS para acceder al cubo vinculado.

Se pueden incluir en los cubos celdas calculadas, este tipo de celdas se calcula en tiempo de ejecución. La celda calculada o celdas calculadas pueden involucrar mediante una expresión multidimensional (MDX) a una o varias celdas, incluyendo cálculos condicionados por otras celdas.

MDX, Multidimensional Expressions (expresiones multidimensionales), es una sintaxis que admite la definición, manipulación de objetos y datos multidimensionales. MDX es similar en muchos aspectos a la sintaxis de SQL (Structured Query Language, lenguaje de consulta estructurado), pero no es una extensión del lenguaje SQL, de hecho, SQL puede ofrecer parte de la funcionalidad que ofrece MDX, aunque no con tanta eficacia ni de manera tan intuitiva. La principal característica es que una sentencia SQL devuelve un resultado basado en dos dimensiones, mientras que una consulta MDX puede incluir varias dimensiones, devuelve un resultado multidimensional, que es un corte o rebanada del cubo encuestado.

AS incluye la tecnología de minería de datos de manera que pueda ser utilizada para descubrir información implícita en los cubos OLAP, para esto proporciona algoritmos potentes de entrenamiento de modelos de minería de datos con datos procedentes de los cubos, algoritmos de clasificación como *Microsoft® Decision Trees,* entre otras cosas.

La herramienta permite definir acciones sobre un cubo seleccionado o una parte del cubo incluidas las dimensiones, los niveles, los miembros y las celdas. La operación puede iniciar una aplicación que utiliza como parámetro el elemento seleccionado o recuperar información del elemento seleccionado. Esto brinda gran flexibilidad al usuario de conectar los datos recuperados con otra aplicación de la organización.

El servicio PivotTable es utilizado para que las aplicaciones puedan interactuar con los datos multidimensionales. Este provee métodos para el análisis online y offline. Con el servicio se pueden desarrollar aplicaciones usando variedad de técnicas y ambientes. Por ejemplo usando Microsoft ActiveX® Data Objects (Multidimensional, ADO MD) en Microsoft Visual Basic®, o en Active Server Pages (ASP) para un sitio Web. Varias aplicaciones de Microsoft traen incorporadas el PivotTable como por ejemplo Excel 2000, con el cual se pueden manipular los datos dinámicamente, consultarlos en forma gráfica, etc.

4.7 Resumen.

La tecnología OLAP se muestra como una poderosa herramienta a los analistas, brindándoles facilidades nunca antes mostradas, haciendo fácil realizar consultas administrativas y de análisis sobre un gran volumen de información. Como consecuencia de esta variedad de comodidades, distintivos de esta tecnología, ha tenido una aceptación acelerada en el mercado. Por esto es de gran importancia su estudio y aplicación en los procesos en los cuales se pueda usar para lograr su eficiente automatización. Teniendo en cuenta esto se ha realizado este estudio para adentrarnos en estas nuevas soluciones que se que se imponen más cada día. En este capítulo se ha presentado una introducción sobre los aspectos básicos a conocer para utilizar cualquier herramienta OLAP.

Conclusiones

Al terminar el tiempo para el desarrollo del trabajo se han resuelto de manera satisfactoria los objetivos propuestos.

1) El algoritmo de búsqueda fonética para el idioma español fue implementado y probado de manera satisfactoria utilizando la BD Persona del centro, obteniéndose el resultado esperado, de ampliar los objetos coincidentes en una búsqueda de nombres con similitud de sonido. Brindándole al usuario una variante más flexible al buscar un individuo.

2) Se realizó un estudio sobre posibles variantes de arquitectura para el Centro de Datos basado en la potencia y posibilidades de los gestores de Base de Datos y los Sistemas Operativos principalmente SQL Server 2000 y versiones de Windows respectivamente, arribándose a una propuesta que cumple con los requisitos pedidos. La implantación de la infraestructura analizada se traduce en una alta disponibilidad a los datos del centro, indispensables para numerosas aplicaciones.

3) Se describieron los aspectos básicos a tener en cuenta sobre las herramientas OLAP y los Cubos Multidimensionales, que constituye su principal objeto. Así como la introducción de las características principales de una herramienta en particular, Analysis Services de SQL Server 2000. La introducción de estas técnicas brinda una forma eficiente y rápida, a los directivos del centro, de analizar grandes volúmenes de información.

La aplicación de los temas estudiados implicarán en resumen, un mejor servicio al ciudadano UCI.

Recomendaciones

Hechas las conclusiones del trabajo, se recomienda:

1) Implantar la arquitectura de clúster propuesta para dar solución a los problemas de disponibilidad en el futuro Centro de Datos de la UCI.

2) Realizar un estudio más profundo sobre la tecnología de cluster basada en el sistema operativo Linux y el gestor de BD Oracle.

3) Realizar estudios y comparación de otras Herramientas OLAP y escoger la más adecuada.

Referencias Bibliográficas

- [1] C. J. DATE, *Sistemas de Bases de Datos 1era Parte.*
- [2] Lesk M., *How much information is there in the world? ,* http://www.lesk.com/mlesk/ksg97/ksg.html. (20/04/2004)
- [3] Schneider, G.M. and Gersting, J.L. *An invitation to computer science*, 2nd ed. PWS Publishing, Pacific Grove, CA, 1998.
- [4] Claudio Casares , http://www.programacion.com/bbdd/tutorial/warehouse/2/ . (05/05/2004)
- [5] Presentación de tesis, Universidad de las Américas-Puebla, http://www.udlap.mx/~tesis/lis/garcia_j_p/. (10/04/2004)
- [6] Sistemas de Base de Datos, http://usuarios.lycos.es/cursosgbd/UD2.htm. (12/05/2004)
- [7] Megaservidores SQL Server: escalabilidad, disponibilidad y capacidad de administración, http://www.microsoft.com/spain/technet/recursos/articulos/welcome3.asp?opcion=1006031#Indicadores%20de%20escalabilidad. (18/05/2004)
- [FON01] Fonética y fonología http://www.lablaa.org/ayudadetareas/espanol/espa16.htm, (02/06/2004)

- Lawrence Philips' Metaphone Algorithm, http://aspell.sourceforge.net/metaphone/. (20/04/2004)
- Algoritmo Soundex, http://www.configuracionesintegrales.com/miguele/soundex.asp?articulo=220. (20/04/2004)
- Dev Articles, http://www.devarticles.com/c/a/Development-Cycles/Tame-the-Beast-by-Matching-Similar-Strings/.(10/05/2004)
- Top Ten TPC-C by Price/Performance, http://www.tpc.org/tpcc/results/tpcc_price_perf_results.asp .(25/05/2004)
- Controlador RAID expandible PowerEdge 4/Di de Dell™ Guía del usuario del controlador PERC4/Di, http://support.ap.dell.com/docs/storage/perc4di/sp/ug/intro.htm#1105329. (28/05/2004)
- Alta disponibilidad para Linux, http://es.tldp.org/Presentaciones/200103hispalinux/paredes/html/x135.html. (2/06/2004)
- Linux Virtual Server Project, http://www.linuxvirtualserver.org/. (2/06/2004)
- Products Designed for Microsoft Windows – Windows Catalog and HCL, http://www.microsoft.com/whdc/hcl/default.mspx. (25/05/2004)
- Libros en pantalla de SQL Server 2000.
- Enlaces de interés relacionados con Tecnologías de Clustering http://www.certificacionmcse.com/mundow2000/clustering.htm .(10/05/2004)
- Alta Plana, Online Analytical Processing http://altaplana.com/olap/olap.documentation.html (15/05/2004)