Universidad de las Ciencias Informáticas

Facultad 9

Título: Estrategia para la replicación de datos espaciales en Sistemas de Información Geográfica.

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Autor:

Yasmany Zapata Pérez

Tutora:

__Ing. Annabell Schelton Lima

Co-tutora:

__Lic. Yaima Fernández Segredo

Consultante

_Ing. Dania Mora Valdes____

Asesor:

__Lic. Yinet Marzo Manue ___

Ciudad de la Habana, diciembre 2008.

"Año 50 de la Revolución"

DEDICATORIA

AGRADECIMIENTOS

DATOS DE CONTACTO

Síntesis de la Tutora
Profesión: Desarrolladora del Módulo de Catalogo de GSIG
Categoría docente: Instructor Recién graduado
Años de graduada: 2008

Síntesis de la Co-tutora
Profesión: Profesor
Categoría docente: Instructor Recién graduado
Años de graduada: 2008

Síntesis del Consultante
Profesión: Profesor
Categoría docente: Instructor Recién graduado
Años de graduada: 2008

Síntesis de la Asesora
Profesión: Especialista funcional, consultora
Categoría docente: Instructor Recién graduado
Años de graduada: 2008

OPINIONES Y AVALES

OPINIÓN DEL TUTOR

RESUMEN

El presente trabajo de investigación, titulado: “Estrategia para la replicación de datos espaciales en Sistemas de Información Geográfica”, se realiza como alternativa de solución a la inexistencia de una estrategia para la replicación de datos espaciales, que permita hacer esta operación sólo a informaciones que son relevantes para el cliente siempre y cuando no afecte la integridad y seguridad del sistema que provee dicha información. Luego de desarrollarse un estudio referente a los entornos y modelos de replicación de datos, se determinó que el maestro/esclavo y el asíncrono, respectivamente, son los ideales a utilizar; asimismo se estableció el empleo de un mecanismo de replicación de datos propio surgido a raíz de la investigación de los demás mecanismos existentes. El uso de la herramienta “Magic@ Data Replication eXtensible Solution” permitió fortalecer esta estrategia, logrando acceder a los datos socioeconómicos relevantes y relacionarlos con los espaciales sin modificar las bases de datos originales.

PALABRAS CLAVE

Datos espaciales, minería de datos, replicación de datos.

DATOS EN INGLÉS

TABLA DE CONTENIDO

TABLA DE CONTENIDO

TABLA DE CONTENIDO

INTRODUCCIÓN

A partir del aumento de velocidad, memoria y disminución de tamaño de las computadoras, las tecnologías de las comunicaciones y en particular la tecnología informática han experimentado enormes progresos, surgiendo un concepto importante, las Tecnologías de la Información y las Comunicaciones (TIC). Así, estas "son un conjunto de servicios, redes, software y dispositivos que tienen como fin la mejora de la calidad de vida de las personas dentro de un entorno, y que se integran a un sistema de información interconectado y complementario, y están destinada a optimizar la comunicación humana.". Actualmente los que interactúan con esta tecnología, de una forma u otra, reconocen la importancia que reviste el empleo de las TIC, siendo imposible por tanto perfeccionar nuestra labor sin su aplicación.

En Cuba, las TIC se aplican en todas las esferas de la vida, en el sector de la salud, los servicios, la educación, las investigaciones y la gestión económica.

La Universidad de las Ciencias Informáticas (UCI), se yergue como principal impulsor de la aplicación de estas tecnologías en el país, pues es un centro destinado al estudio y desarrollo de la tecnología informática.

Entre los sistemas que se desarrollan hoy en dicha universidad, se distinguen por su impacto en el ámbito social y económico, los Sistemas de Información Geográfica, conocidos como SIG.

Se puede decir que un SIG es un conjunto de métodos, herramientas y datos que están diseñados para actuar coordinada y lógicamente para capturar, almacenar, analizar, transformar y presentar toda la información geográfica y de sus atributos con el fin de satisfacer múltiples propósitos.

En la UCI, el grupo de desarrollo de este tipo de sistemas, se ha dado a la tarea de desarrollar soluciones para diferentes entidades, cada una con características particulares, pero siempre partiendo de la necesidad común de contar con un SIG que apoye, entre otros procesos, la toma de decisiones y facilite el análisis de la información espacial. Para cumplir con esta misión, el Grupo de

INTRODUCCIÓN

Desarrollo de Sistema de Información Geográfica (conocido por sus siglas, como GSIG) debe trazarse una estrategia de desarrollo que le permita de forma paralela desarrollar varias soluciones, teniendo en cuenta las especificidades de cada cliente. Esto se debe a que en estos sistemas se manejan grandes volúmenes de información y de carácter variado además, pues se combinan datos espaciales con datos socioeconómicos, y por consiguiente resulta una tarea difícil encontrar una solución genérica para el almacenamiento y control de la información.

Otro aspecto a considerar es que cada cliente generalmente posee una estructura propia para almacenar sus datos socioeconómicos, estructura que no se debe modificar en función de relacionarla con la representación espacial de cada objeto, elemento imprescindible en la realización de un SIG. Por tanto, puede definirse como la **Situación Problemática** de la investigación *la necesidad que hoy presenta el GSIG de encontrar un mecanismo factible para que de manera independiente a la entidad a la que se le esté personalizando una solución, se pueda acceder tanto al conjunto de datos espaciales como a los datos socioeconómicos que sean de mayor relevancia para el cliente, garantizando que no se modifiquen las bases de datos originales y que la calidad del sistema no se vea afectada.*

A partir del análisis de la situación problemática de la investigación se define como el **Problema a resolver** la siguiente interrogante**:**

¿Cómo lograr una estrategia segura, eficiente y efectiva para que GSIG pueda acceder a los datos socioeconómicos relevantes y relacionarlos con los datos espaciales sin modificar las bases de datos originales?

Para llevar a cabo la investigación es necesario estudiar el *Proceso de replicación y minería de datos* que constituye el **Objeto de estudio**, para luego incidir directamente en *Los mecanismos aplicados en la replicación de datos para SIG,* que representa el **Campo de acción.**

La meta a alcanzar en el presente trabajo para solucionar el problema queda sujeta al siguiente **Objetivo general**: *Establecer una estrategia eficiente para las réplicas de datos espaciales en SIG.*

La **idea a defender** del trabajo es que con la puesta en práctica de la estrategia que se propone en la investigación se lograría un proceso de réplica de datos para SIG bien definido y eficiente.

Para cumplir con el objetivo trazado en esta investigación, es necesario llevar a cabo un conjunto de **tareas investigativas** que permitan, de manera sistemática y creciente, ir avanzando en el desarrollo de la investigación para lograr dar solución al problema científico que le dio origen a la misma:

- Caracterización de modelos de replicación de datos.
- Proposición de modelos de replicación de datos.
- Caracterización de métodos de minería de datos, para seleccionar la información relevante.
- Proposición de un método de minería de datos a usar.
- Caracterización de los mecanismos existentes para la replicación de datos.
- Comparación de los mecanismos para replicación de datos existentes.
- Selección de un mecanismo factible para la replicación de datos espaciales en los sistemas de información geográfica.
- Confección de una arquitectura conceptual para la replicación de datos espaciales en los sistemas de información geográfica.
- Documentación de la estrategia definida en función de la socialización de la misma.

Métodos de investigación científica:

Mediante los métodos de investigación científica es posible un estudio detallado de los orígenes de la replicación, así como sus características, campos de desarrollo.

Métodos teóricos:

Histórico-lógico: En la primera fase de la investigación se desarrolla un estudio del estado del arte de la problemática analizada, revisando de forma crítica cada uno de los documentos lo que ha permitido resaltar la importancia de las diferentes estrategias para la replicación de datos que se llevan a cabo en la actualidad en el mundo, así como las ventajas y desventajas de los mecanismos de replicación de datos, además de los modelos asociados a la problemática investigativa. Asimismo, se permitió conocer detalladamente la trayectoria y desarrollo de aplicaciones SIG teniendo en cuenta sus precedentes históricos, las investigaciones realizadas y los resultados obtenidos por otros autores al efectuar dichas investigaciones.

Análisis-Sintético:

- **Análisis**: Este método por su particularidad es de gran utilidad, puesto que permite desglosar el fenómeno, dando la posibilidad del estudio pormenorizado de la replicación de datos en cada una de sus partes y componentes.

- **Síntesis**: Con el estudio de los SIG, este método brinda la posibilidad de integrar todo el conocimiento adquirido sobre estos durante el análisis, llegando así a descubrir sus características generales.

Métodos Empíricos:

Entrevista: A través de esta técnica se llegó a adquirir conocimientos e información acerca de los procesos de replicación de datos espaciales, ayudando a la toma de decisión y descubrimiento sobre herramientas de replicación a utilizar.

Estructura de la investigación:

- **En el capítulo 1:** "Fundamentación teórica", se describe todo lo concerniente al objeto de estudio, así como los términos utilizados y el análisis de distintas documentaciones existentes para la replicación de datos espaciales en Sistemas de Información Geográfica.

- **En el capítulo 2:** "Construcción de la Solución", se brinda la propuesta de los mecanismos para la replicación de datos espaciales en sistemas de información geográfica.

- **En el capítulo 3:** "Propuesta y Documentación de la Solución", se presenta la propuesta de documentación mediante los mecanismos elegidos en el capitulo anterior.

CAPÍTULO 1: “Fundamentación Teórica”.

1.1 Introducción

Este capítulo contiene las definiciones y conceptos que serán de utilidad para el entendimiento de la investigación. Trata el tema del surgimiento y desarrollo de los SIG hasta el presente y su aplicación en Cuba, el análisis de soluciones existentes y se abordará con mayor profundidad la situación problemática que da origen a este trabajo.

1.2 Conceptos asociados al dominio del problema

Para la correcta comprensión del trabajo investigativo, es necesario especificar el significado de algunos conceptos, que son, conducentes y esenciales objetivos para lograr desarrollar la estrategia definitiva.

1.2.1 Sistema de Información Geográfica (SIG)

Los SIG poseen gran importancia tanto en la esfera social como económica, atendiendo además a que la solución que se propone en esta investigación va encaminada a este tipo de sistemas, se hace imprescindible abordar con profundidad qué es un SIG, y cuáles son sus principales beneficios.

Un **SIG** puede entenderse como un poderoso conjunto de herramientas para adquirir, almacenar, recuperar a voluntad, transformar y desplegar datos espaciales para determinados propósitos.

Técnicamente se puede definir un SIG como una tecnología de manejo de información geográfica formada por equipos electrónicos (hardware) programados adecuadamente (software) que permiten manejar una serie de datos espaciales (información geográfica, datos geográficos) y realizar análisis complejos con estos siguiendo los criterios impuestos por el equipo científico (personal o equipo humano).

Son, por tanto, cuatro los elementos constitutivos de un sistema de estas características:

1. Hardware.
2. Software.
3. Datos geográficos.
4. Equipo humano.

Podrían citarse otras definiciones, pero en esencia pueden concretarse como: sistemas con la capacidad de manipular datos espaciales, brindándonos de esta manera una herramienta que permite visualizar y analizar la información de forma versátil e intuitiva, agilizando la tan importante toma de decisiones.

1.2.2 Datos Espaciales

La replicación que se propone a realizar debe incluir los datos socioeconómicos y también los datos espaciales significativos, por tanto es de gran importancia definir su concepto y conocer su escala en esta investigación:

Los **Datos Espaciales** son el componente fundamental de cada proyecto o aplicación SIG. Contienen las ubicaciones y formas de características cartográficas. Son también conocidos como Datos Cartográficos Digitales, el tipo de datos necesarios para crear mapas y estudiar relaciones espaciales.

Dentro de su contexto, almacenan informaciones sobre la localización y las formas de un objeto geográfico y las relaciones entre ellos, normalmente con coordenadas y topología. Refieren a entidades o fenómenos que cumplen los siguientes principios básicos:

- **Posición absoluta**: sobre un sistema de coordenadas (x, y, z).
- **Posición relativa**: frente a otros elementos del paisaje (topología: incluido, adyacente, cruzado, etc.)
- **Figura geométrica** que representan (punto, línea, polígono)
- **Atributos** que describen (características del elemento o fenómeno)

[1]

1.2.3 Bases de datos (Database, BD)

Una **base de datos** es una colección de información organizada de forma que un programa de ordenador pueda seleccionar rápidamente los fragmentos de datos que necesite. Es un sistema de archivos electrónico, para su posterior uso en la recolección y manejo de datos.

Las bases de datos tradicionales se organizan por campos, registros y archivos. Un campo es una pieza única de información, un registro es un sistema completo de campos; y un archivo es una colección de registros. Por ejemplo, una guía de teléfono es análoga a un archivo. Contiene una lista de registros, cada uno de los cuales consiste en tres campos: nombre, dirección, y número de teléfono.

A veces se utiliza DB, de *database* en inglés, para referirse a las bases de datos. [2]

1.2.4 Datos Socioeconómicos

Los **datos socioeconómicos** son aquellos datos en cuyo contexto se relacionan los datos de personas, contando con sus propiedades y atributos que pueden ser: nombre, apellido, identificación, salario, etc.

Son los datos que en el marco de la investigación se relacionan con los espaciales para agregar información útil a un objeto, con el objetivo de ser consultada.

Los datos espaciales por sí solos permiten la localización de los objetos, sin embargo cuando se relacionan con datos socioeconómicos, las aplicaciones se robustecen en gran medida, pues brindan una mayor cantidad de información, además variada y personalizada al entorno que se modele, lo que posibilita tomar decisiones con mayor facilidad.

1.3 Proceso de replicación y minería de datos

1.3.1 Descripción General

En los últimos años, con el auge del campo de la informática, y su aplicación del tratamiento automático de la información utilizando dispositivos electrónicos y aplicaciones computacionales, se han creado sistemas cada vez más complejos que constan de una multitud de elementos interconectados a través de redes de comunicaciones rápidas.

La configuración óptima de cada uno de estos elementos para obtener el máximo rendimiento, depende de un gran número de factores como son la carga del entorno, su tipo y el hardware sobre el que se ejecuta, entre otros. Esta diversidad de parámetros de configuración hace que la administración sea cada vez más complicada y que requiera de administradores con mucha experiencia y dedicación a estas tareas. Por esta razón ha surgido la necesidad de crear sistemas que sean capaces de adaptarse de forma automática a los cambios en el entorno que les rodea sin necesidad de intervención humana y que al mismo tiempo exhiban un alto rendimiento y calidad de servicio.

Estos sistemas adaptables dinámicamente interaccionan con su entorno para detectar los cambios en el mismo, analizan la información obtenida para generar configuraciones óptimas a las condiciones actuales del entorno y modifican su configuración para adaptarse a estos cambios. El objetivo de estos sistemas es maximizar el rendimiento, de acuerdo a las métricas de rendimiento de interés, y minimizar las pérdidas.

Para darle una solución factible y poco riesgosa a estas problemáticas, comienza la utilización de la replicación, surgiendo como un término a la línea de investigación a finales de los años 80, con el propósito de buscar una solución al problema de descubrimiento de bases de datos. Juega un papel muy importante como forma de aumentar el rendimiento y las ganancias en materia de información.

Hasta ahora la replicación se implementaba dentro de la propia base de datos, lo que obligaba a realizarle modificaciones y estas no siempre eran posibles, por diferentes razones y aspectos. Uno de los factores es el temor a la pérdida o modificación de la información, tanto sensible como de orden muy significativo. En otros casos se planteaba la necesidad de sólo replicar aquellos datos que fuesen significativos a la necesidad. En contenido, es una técnica que permite copiar y distribuir idénticamente las tablas de una BD en otras múltiples ubicadas en diferentes nodos de la red. Todo ello hace posible que los datos correctos estén siempre disponibles en el momento y lugar necesarios, poseyendo las características de: Efectividad, Alta Disponibilidad, Tolerancia a Fallos y Coordinación.

Contando con grandes cantidades de datos, y las dificultades para encontrar estos datos ocultos y extraerlos, llegando hasta ser crítica, en las BD surge la tecnología de Minería de Datos que ha estado bajo desarrollo por décadas, en áreas de investigación como estadísticas, inteligencia artificial y aprendizaje de máquinas. El alcance de esta tecnología, deriva de las semejanzas entre buscar información valiosa en grandes BD y buscar una aguja en un pajar. Para su realización, se requiere examinar una inmensa cantidad de material, o investigar inteligentemente hasta encontrar exactamente dónde residen los valores. Localizar y obtener la información que se quiere, sólo la necesaria.

1.3.2 Situación Problemática

Los SIG poseen varios beneficios, constituyen una excelente solución para el manejo de datos en campos como la estadística, construcciones y la planificación del uso del suelo. La UCI con la creación del GSIG, garantiza el desarrollo de estos sistemas en tiempos en los que su uso se ha incrementado e insertado en varias esferas socioeconómicas.

En la actualidad este Grupo está desarrollando una plataforma modular que contenga aquellas funcionalidades que le son comunes a cualquier SIG, sobre tecnología libre y con capacidad de personalizarse a entornos específicos.

El proceso de personalización es complejo, por lo que GSIG hoy busca soluciones genéricas que le permitan una adaptación rápida y sencilla. Para ello se capacitan nuevos equipos de trabajo que se especialicen en esta tarea, y se definen líneas de investigación para garantizar que los campos principales dentro de la línea SIG estén cubiertos.

Una de las problemáticas que ha sido identificada dentro del proceso de personalizar la plataforma a un entorno determinado, es la de manejar los datos socioeconómicos y espaciales. Esto ocurre porque generalmente cada cliente posee una BD propia donde almacena los datos socioeconómicos de su entidad, y sobre la que corren otras aplicaciones. Por tanto el hecho de relacionar los datos que sean relevantes de ese conjunto para el cliente, con los datos espaciales sin atentar contra la seguridad de otras aplicaciones que se beneficien de la BD maestra, y además garantizando que no se modifiquen las fuentes originales, y que la calidad del sistema no se vea afectada, es una problemática compleja que debe solucionarse mediante una estrategia general que permita guiar a los desarrolladores en el proceso de personalizar.

Brindar esta solución es el objetivo que se persigue con la presente investigación.

1.4 Análisis de otras soluciones existentes

A escala mundial se han desarrollado soluciones a la problemática que da origen a esta investigación mediante la replicación de datos, aplicándola según sus necesidades.

En el presente epígrafe se hace un análisis de dichas soluciones para intentar buscar apoyo en alguna de ellas, que sirva de base para proponer la solución de esta investigación, o comprobar si existe alguna semejante que permita enfrentar al problema científico planteado. Por ejemplo:

Hitachi Data Systems, tiene una amplia y probada trayectoria en soluciones que les ofrecen a sus clientes una cantera completa de productos, servicios y soluciones para la continuidad de negocios.

Esta compañía estuvo presentando una solución de replicación heterogénea para entornos virtuales, llamada: Hitachi Storage Replication Adapter, que apunta a disponibilidad en automático y de replicación continua, remota y local, en tiempo real, para garantizar flexibilidad y solidez en la recuperación de datos, alta disponibilidad, basándose principalmente en la recuperación de desastres para entornos virtuales VMware. A medida que en el mercado se acelera la adopción de la virtualización de servidores, las organizaciones buscan la forma de reducir el riesgo, el costo y la complejidad que resultan de aplicar los métodos de recuperación de desastres tradicionales a cargas de trabajo virtualizadas. Los clientes se benefician con una mayor protección de datos y una menor exposición a riesgos, lo cual mejora la flexibilidad y robustez operativa.

Por otra parte **Microsoft SQL Server 7**, perteneciente a la multi-gigantesca compañía Microsoft Corporation, propone un modelo de replicación de datos para darle soluciones a diferentes problemas, empleando tres modelos de replicación. Si el problema reside en cómo hacer para que el publicador replique datos tal como están en la base de datos en un momento dado, hace uso del Snapshot Replication. El funcionamiento es sencillo: el publicador simplemente envía una réplica de todos los datos hacia los suscriptores, en vez de solamente enviar los datos que fueron alterados desde el último snapshot realizado. Otra de las soluciones es realizando un monitoreo de los cambios a los datos que son realizados en el publicador (inserciones, borrados y modificaciones de transacciones que dieron COMMIT), como parte del método: **Transactional Replication**. Este mecanismo garantiza una consistencia transaccional en sentido laxo: todos los suscriptores tarde o temprano reciben los datos como si se hubiesen alterado en un único sitio. Por último es utilizado el método Merge **Replication**, para solucionar los problemas referentes a permitir que los diferentes servidores actúan con alto nivel de independencia y desconectados entre sí. La replicación Snapshot y Transactional se basan en un modelo de replicación en una sola dirección, desde un único publicador hacia los suscriptores. Mientras que Merge mueve y reconcilia los cambios a los datos, ocurridos después de la inicialización de la replicación. Los datos se pueden mover en ambas direcciones o en una sola. [3]

Otra de las soluciones a tener en cuenta en el análisis es la que ofrece **Cybertec Schönig & Schönig GmbH**, que es una compañía que posee una gran gama de servicios para el mundo de las Bases de Datos Open Source. Entre ellas se encuentra una completa y altamente balanceada

solución de replicación multi-master para PostgreSQL. En muchos casos la replicación asincrónica no es suficiente para modelar ciertos problemas, por lo que ofrece una solución sincrónica Multi-Master de replicación para PostgreSQL llamada **Cybercluster**. Esta solución es aplicada a aquellos entornos donde un clúster de Bases de datos es consistente en cada punto del tiempo, basándose en una arquitectura shared-nothing conveniente para la replicación sincrónica Multi-Master. [4]

Otro ejemplo de este tipo de soluciones es **HIT Software**. La línea de productos HIT Software es una óptima solución que incluye un conjunto de herramientas de integración de datos que operan en estándares abiertos. Con DBMoto, Hit Software brinda una replicación de datos en tiempo real y una integración de datos. DBMoto su sustenta en realizar una actualización y replicación de los datos en tiempo real en el servidor de la empresa del cliente y las estaciones de trabajo que requieran de replicación. Las principales bases de datos como IBM DB2 UDB (incluidos los de System i; i5/iSeries/AS400 y zOS), Oracle, Microsoft SQL Server, Sybase ASE, SQL Anywhere, Cloudscape, MySQL, Informix, Ingres, PostgreSQL, Microsoft Access, Gupta SQLBase, Firebird y Solid son soportadas para estos fines. Para su funcionamiento y actividad, se experimenta con tres modos, los cuales son: Refrescar, Espejamiento y Sincronización, tales modos, son también conocidos en inglés: Refresh, Mirroring y Synchronization, respectivamente. [5]

En modo Refrescar (refresh), DBMoto, lee los datos, se aplican en el administrador pre-definido las reglas a cumplir y escribe el resultado en la base de datos de destino. En modo Espejamiento (Mirroring), realiza un aumento en la replicación en tiempo real basándose sobre el registro de transacciones (logs). En el modo Sincronización (Synchronization), proporciona la capacidad de mantener las bases de origen y destino sincronizadas.

La solución **Double-Take® de Double-Take® Software** protege los datos permitiendo aplicaciones como Exchange, SQL Server, SharePoint, Oracle y otros, posibilitando a los clientes replicar y proteger la información crítica de la empresa. Algunos de los beneficios de la Replicación de Datos de Double-Take son:

- Protección de Datos en Tiempo Real.

- Replica continuamente a nivel de byte sobre cualquier LAN, WAN o SAN, asegurando que los cambios de los datos sean protegidos y que puedan ser restaurados rápidamente en cualquier momento.
- Agnóstico de Aplicación
- Trabaja con su hardware existente protegiendo aplicaciones como Exchange, Microsoft SQL Server, Oracle, SharePoint, y otros.
- Protección Continua de los datos a través de la Replicación
- Garantiza la continuidad del negocio y una alta disponibilidad utilizando la replicación en tiempo real y restaurando el acceso a los datos en minutos y con la capacidad del failover (teniendo la configuración de un segundo, asumir el cargo, si falla el primero) para obtener un entorno de trabajo íntegro.
- Fácil de instalar y mantener
- Permite a compañías de cualquier tamaño buscar una solución para la protección de sus datos.

  Rentabilidad
- Provee la mejor protección posible al menor costo con un acelerado retorno de inversión – pagándose el mismo en un período de pocos meses. [6]

A pesar de la eficiencia que pueden tener estas soluciones en algunos entornos, no son viables para ser utilizadas por GSIG. Inicialmente se está buscando una solución sobre plataforma libre y no la utilización de software propietario, sobre todo por la utilización de PostgreSQL. El objetivo tampoco es realizar una réplica con el fin de recuperarse ante desastres o como uso de backup, sino ~~de~~ aplicar técnicas de minería de datos para elegir la información relevante llegando a producirse una interacción entre la BD original y la de destino. El hecho de que se necesite una solución para SIG, ya es una condición que estrecha en gran medida el espectro de soluciones posibles, pues estos Sistemas poseen características singulares que los hacen únicos, además de que no se trata de una solución específica a un entorno sino genérica, lo que acentúa la complejidad de la estrategia a desarrollar.

Hasta el momento no se ha encontrado una solución que dé respuesta directa al problema que da origen a esta investigación.

1.5 Conclusiones

El desarrollo del capítulo comprendió los conceptos básicos para el entendimiento y surgimiento de la investigación. Se hizo un análisis más profundo del objeto de estudio, así como de la situación problemática, además se realizó un análisis de las soluciones existentes que sirve como punto de partida para continuar el estudio.

CAPÍTULO 2: “Construcción de la Solución”.

2.1 Introducción

En este capítulo se desarrolla el Análisis, Selección, Comparación de los Modelos, Mecanismos y Diseño de la estrategia, brindándose una propuesta de los modelos y mecanismos para la replicación de datos espaciales en SIG. Una introducción a la minería de datos, sus fases y técnicas constituyen además aspectos esenciales de este capítulo.

2.2 Análisis de los Modelos de Replicación de Datos

La **replicación** es un conjunto de tecnologías destinadas a la copia y distribución de datos y objetos de base de datos, desde una a otra, para luego sincronizarlas y mantener su coherencia. Este proceso permite distribuir datos entre diferentes ubicaciones y entre usuarios remotos o móviles mediante redes locales y de área extensa, conexiones de acceso telefónico, conexiones inalámbricas e Internet. [7]

Replicación también se entiende por un mecanismo que permite mantener copias actualizadas automáticamente (si fuese el caso) de los datos de un servidor de bases de datos en otros. Un proceso de compartir información a fin de garantizar la coherencia entre los recursos redundantes, tales como programas informáticos y componentes de hardware, para mejorar la fiabilidad, la tolerancia a fallos, o la accesibilidad. [8]

Existen distintos tipos de réplicas: la **activa** y **pasiva**. La primera se realiza mediante la transformación de la misma petición en cada réplica. La segunda**,** en cada solicitud se tramita en una única réplica de su estado y luego se transfiere a las demás réplicas.

Para efectuar la replicación, esta se lleva a cabo por medio de modelos, los cuales aportan una estrategia a seguir y una guía, para el desarrollo organizado de la replicación. Un **modelo** es un

arquetipo digno de ser imitado que se toma como pauta a seguir [9]. Útil para ser imitado, reproducido o copiado.

Los **modelos de replicación** son la forma en que se han llevado a la práctica el enfoque de replicación de datos por distintas empresas y entidades que los han adoptado. Se describen como un mecanismo que permite mantener copias actualizadas automáticamente (si fuese el caso) de los datos de un servidor de bases de datos en otros, los mismos servirán como base de la investigación a partir de los elementos más importantes identificados en cada uno de ellos. Existen dos modelos:

- Sincrónico
- Asincrónico

Encontrándose asociados a ellos dos entornos:

- Multi-Maestro
- Maestro-Esclavo

Estos modelos y entornos, son utilizados en pares y asociativos, es decir, el modelo sincrónico o el asincrónico, con uno de los dos entornos: multi-maestro o el maestro-esclavo. Esencialmente surge el término de **modelo de réplica** para sincrónico y asincrónico, aplicados a los **entornos de réplica** para multi-maestro y maestro-esclavo.

2.3 Características de los modelos de replicación

Analizando los modelos y entornos de réplica, en la fig. 2.1, se muestra un Esquema de Réplica, detallándose las técnicas de replicación, teniéndose una breve representación de los Entornos, Modelos de Distribución (modelos de replicación), Tecnologías y Problemas que surgen en cada caso. Cada entorno de replicación (Maestro/Esclavo, Multi/Maestro) puede asociarse a uno de los modelos de distribución (Sincrónica, Asincrónica), y estos a su vez tienen asociado una o diferentes tecnologías (configuración en dos fases, descarga/carga, instantánea, disparadores/reglas bases de la replicación). Poseyendo algunos inconvenientes (redes y sistemas no disponibles, impuntual,

transacciones inconsistentes, administración-funcionamiento de la sobrecarga), lo que dificulta la toma de decisión sobre cuál modelo utilizar.

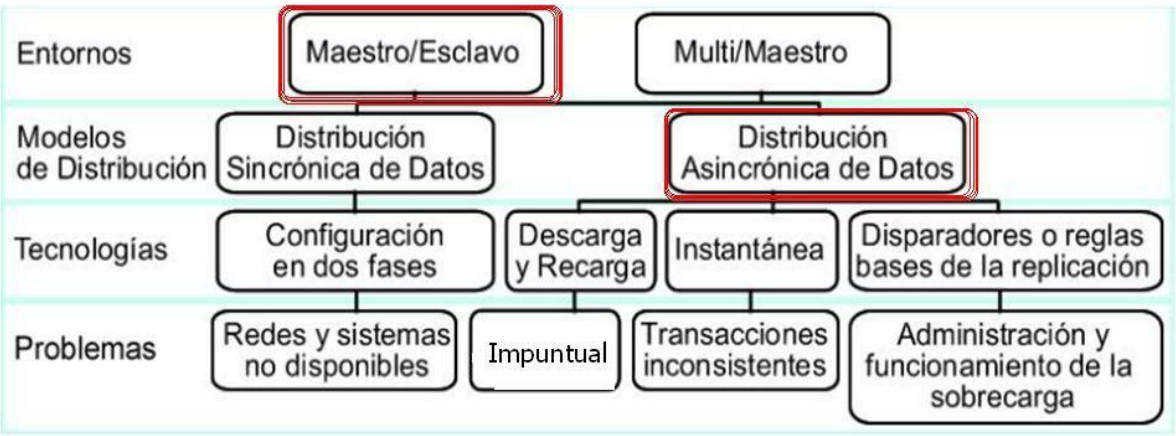

Fig. 2.1 Esquema de replicación

2.3.1 Modelos de replicación

2.3.1.1 Sincrónico

El término sincrónico, proviene de sincronía, que se entiende como un término que se refiere a coincidencia en el tiempo o simultaneidad de hechos o fenómenos. Otro significado que puede atribuírsele es el de ser simultáneo, que ocurre o se desarrolla a la vez que otra cosa.

El **modelo de replicación sincrónico**, también conocido como la réplica en tiempo real, surge a mediados de los años 80 y aplica cualquier cambio o ejecuta cualquier procedimiento reproducido en todos los sitios que participan en el ambiente de réplica como parte de una sola transacción. Si el procedimiento falla en cualquier sitio, entonces la transacción entera se anula. La réplica sincrónica asegura la consistencia de datos en todos los sitios en tiempo real.

Algunas características que presenta este modelo son:

- Actualiza “almacenes” de datos al mismo tiempo. Cada transacción solamente es aceptada si todos los sistemas implicados en la réplica están conectados y listos para recibirla, Si alguno falla, todo el proceso es anulado, como se muestra en la figura 2.2
- Muy fiable e ideal para recuperarse ante desastres.
- Obviamente: alto impacto en la red. Poco escalable y caro

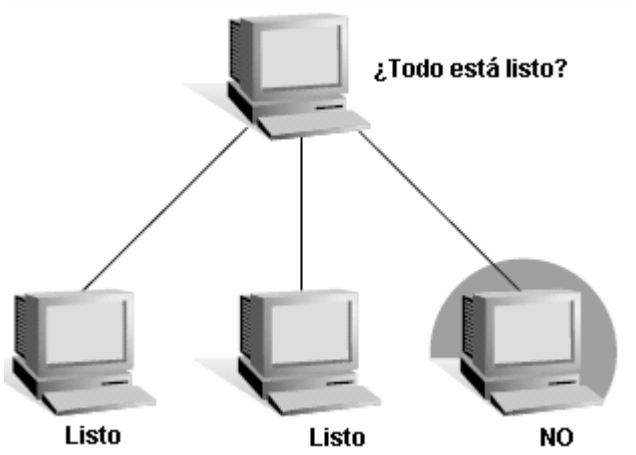

Figura 2.2 Confirmación de dos fases.
Se anula el proceso por no confirmar la última fase.

2.3.1.2 Asincrónico

La replicación **asincrónica**, a menudo llamada almacena-y-reenvía o Descarga y Recarga, captura cualquier cambio local, los almacena en una cola y a intervalos regulares, propaga y aplica estos cambios en sitios remotos. Con esta forma de réplica, hay un período de tiempo antes de que todos los sitios alcancen la convergencia de datos.

Este modelo consiste también en hacer un volcado de los datos, copiar la salva para un dispositivo de almacenamiento para luego distribuir la salva por los demás servidores. Esta técnica presenta el

inconveniente de que en la mayoría de las ocasiones se consultan datos que tienen semanas de desactualización, además de que el proceso se realiza de forma manual.

Este modelo presenta otras características como son:

- Las escrituras se hacen en un “maestro” y con el tiempo se propagan a varios “esclavos”.
- Económico, escalable y flexible.
- Mayor probabilidad de pérdida de datos.

Se puede combinar perfectamente con los entornos de réplica, de acuerdo a los intereses de quien lo utilice, teniendo en cuenta los problemas que surgen una vez compuestos. Estos no dependen explícitamente del entorno, sino del modelo como tal, en este caso el asincrónico, llegando a surgir más inconvenientes que en el sincrónico.

2.3.2 Entorno de Replicación

Básicamente existen dos tipos o entornos de réplicas: el par-a-par o Multi-Maestro (master-master) y el de sólo lectura o maestro-esclavo (master-slave). En el siguiente epígrafe se explican los diferentes entornos.

2.3.2.1 Multi-Maestro

El entorno Multi-Maestro, también llamado par-a-par o la réplica de camino de n, permite múltiples sitios, actuando como pares iguales. Cada sitio en un ambiente de réplica de multi-maestro es un sitio de maestro, y cada sitio se comunica con otros sitios maestros [10]. Esta capacidad tiene también un severo impacto en el desempeño debido a la necesidad de sincronizar los cambios entre los servidores. Lo que también quiere decir la interacción entre servidores en ambos sentidos.

Este movimiento, se realiza de dos o más réplicas de sincronización entre sí, a través de un identificador de transacción. Permite leer/escribir las consultas que se enviarán a múltiples servidores replicados. Esta capacidad también tiene un considerable impacto en el rendimiento debido a la necesidad de sincronizar los cambios entre los maestros.

En la fig. 2.3 se muestra una gráfica de interacción entre dos maestros, donde se refleja la replicación en ambos sentidos.

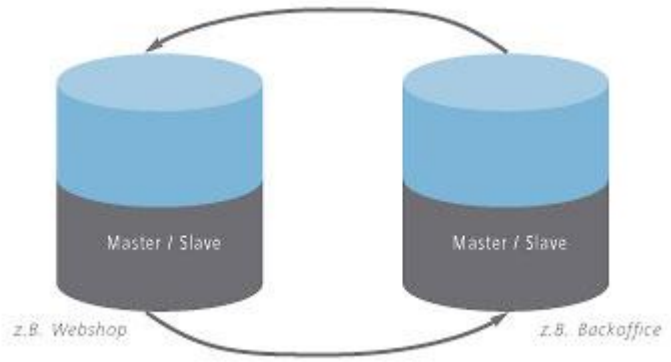

Fig. 2.3

2.3.2.2 Maestro-Esclavo

El entorno Maestro-Esclavo, o de sólo lectura, permite a un solo maestro recibir consultas de lectura/escritura, mientras los esclavos solo pueden aceptar consultas de lectura. Todas las solicitudes de escritura se realizan en el maestro y los esclavos a reproducirse. Esto permite que un solo maestro pueda recibir consultas de lectura/escritura, mientras que los esclavos sólo pueden aceptar la selección de consultas de lectura. Si en algún momento un maestro de réplica ha sido designado para tramitar todas las peticiones, entonces estamos hablando del sistema de copia de seguridad primaria (plan maestro-esclavo) predominante en los grupos de alta disponibilidad. [11]

En la fig. 2.4 se representa una interacción entre el Maestro y sus Esclavos, para llevar a cabo la replicación.

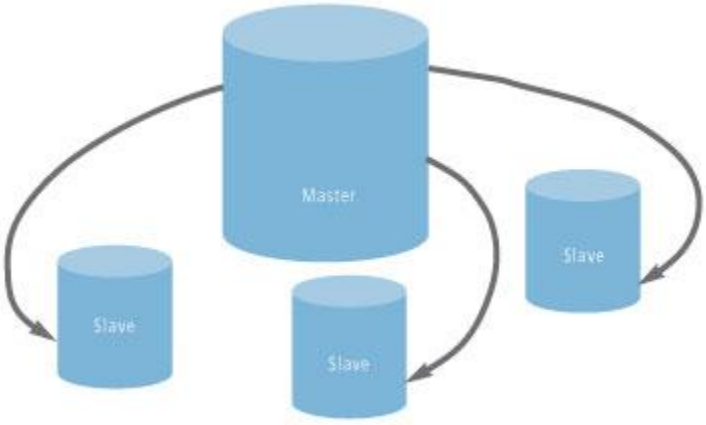

Fig. 2.4

2.4 Comparación de los modelos

En el siguiente epígrafe, se expone una comparación entre los modelos Sincrónico vs. Asincrónico y los entornos de réplica Multi-Maestro vs. Maestro-Esclavo, que ayudará a establecer cuál es el más óptimo utilizar para darle solución al problema en cuestión en esta investigación.

2.4.1 Sincrónico vs. Asincrónico

Con el objetivo de evaluar qué modelo utilizar para la replicación, si el asincrónico o el sincrónico, se confecciona una pequeña comparación de ambos:

En el contexto de la replicación de bases de datos, la definición más común de la replicación **sincrónica**, es decir, que tan pronto como una transacción se confirma, todos los “almacenes de datos” debe tener comprometido también la operación. Esto provoca que si una fase no se confirma, se anula todo el proceso. En los sistemas de replicación **asíncrona**, otros “almacenes de datos” pueden aplicar las transacciones de datos en cualquier punto, por lo tanto, pueden servir diferentes puntos, incluso en conflictos instantáneos de la base de datos.

La **replicación sincrónica** de las bases de datos permite a las transacciones que se realicen simultáneamente en varios servidores, que proporciona un método de copia de seguridad y la seguridad, así como la disponibilidad de datos.

La **replicación asincrónica** puede ser muy exigente en una red, debido al número de variables que deben considerarse. Además de los volúmenes de transacción, otras variables incluyen velocidades de línea, tipo de conexiones, la velocidad y el número de procesadores que participan, la puntualidad de datos y el número de servidores de replicación, también el número de problemas que surgen como las transacciones inconsistentes.

2.4.2 Multi-Maestro vs. Maestro-Esclavo

El modelo **multi-maestro**, posee la peculiar característica de que las réplicas son enfocadas a escribir tanto en un único servidor o almacén, como en los demás, pudiendo escribir y leer en el distribuidor. Mediante el maestro-esclavo las peticiones y asignaciones se realizan desde ambas partes en una sola dirección: los datos se piden desde el esclavo y son dadas por el maestro, en caso de existir uno solo. Si existen varios esclavos o varios maestros, sigue siendo la escritura desde los maestros hacia los esclavos.

2.5 Factores para elegir el modelo de replicación a utilizar

En la elección de un método adecuado para la distribución de los datos, influyen varios factores. Los cuales podemos agruparlos en dos grupos:

1. **Factores relacionados con los requerimientos de la aplicación.**
2. **Factores relacionados con el entorno de red.**
3. **Problemas de acuerdo con la tecnología de replicación utilizada.**
4. **Conflictos de Replicación.**

Dentro de los factores relacionados con los **requerimientos de la aplicación**, los fundamentales son:

- **Autonomía**
- **Consistencia transaccional**
- **Latencia**

La **autonomía** de un sitio, de una aplicación, da la medida de cuánto puede operar el sitio desconectado de la base de datos maestra. Puesto que la replicación puede realizarse en un momento dado y luego, por un espacio de tiempo, no tener acceso a la base de datos original para realizar la solicitud de otra replicación.

La **consistencia transaccional** viene dada por la necesidad de ejecutar o no inmediatamente todas las transacciones que se han ejecutado en el servidor maestro, o si es suficiente con respetar el orden de las mismas. Si se realiza una actualización en el servidor, fuere cual fuere el caso, si es necesario actualizar la BD de destino.

La **latencia** se refiere al momento en que se deben sincronizar las copias de los datos. El instante de acuerdo a la configuración, o según el cliente haya predefinido del proceso en ¿necesitan los datos estar el 100% en sincronía? O si es admisible determinada latencia ¿de qué tamaño es aceptable el atraso?

Entre los **factores relacionados con el entorno de red** están la velocidad de transmisión de datos de la red.

Debe analizarse además la confiabilidad de la red y responder preguntas como:

- ¿Cuán confiable es la red?

Por otra parte en el caso que los servidores SQL no permanezcan todo el día encendidos, como pudiera suceder, deben considerarse los horarios de disponibilidad de cada servidor, el de datos espaciales y los socioeconómicos.

Teniendo en cuenta estos factores, estos sirven de guía en la configuración del ambiente de replicación. Además debe considerar las siguientes preguntas:

- ¿Qué datos se van a publicar?
- ¿Reciben todos los suscriptores todos los datos o sólo subconjuntos de ellos?
- ¿Se deben particionar los datos por sitio?
- ¿Se debe permitir que los esclavos envíen actualizaciones de los datos? Y en caso de permitirlas, ¿cómo deben implementarse?
- ¿Quiénes pueden tener acceso a los datos?

Entre los **factores relacionados con los problemas de acuerdo con la tecnología de replicación** utilizada se encuentran:**:**

- **Redes y sistemas no disponibles** (usando la tecnología de ***configuración en dos fases)***: este inconveniente se presenta en el modelo sincrónico. Por medio de ***configuración en dos fases***, permite la sincronización de datos distribuidos. Cada transacción solamente es aceptada si todos los sistemas implicados en la réplica están conectados y listos para recibirla, si al menos uno falla, todo el proceso es anulado.

Los demás problemas, específicamente en el modelo asincrónico, con el uso de diferentes tecnologías son:

- **Impuntual** (tecnología: ***descarga/carga***):
  - Cuando la realización de la replicación se efectúa mediante dispositivos externos a la red, pudiendo ser una memoria flash USB, que contendría un archivo (dígase un .xml u otra extensión afín) ya previamente descargado ahí, con los datos para ofrecerlos a la base de datos destino, se produce lo que se llama: descarga/carga. Descarga del servidor origen y carga en el de destino. El inconveniente es que en la mayoría de las ocasiones se consultan datos que tienen semanas de des-actualización, además que el proceso se realiza de forma manual.

- **Transacciones inconsistentes** (tecnología: ***Instantánea(Snapshot)***):
  - Mediante la tecnología Instantánea, se replican datos como están en la base de datos en un momento dado. La publicación se puede realizar en forma cronogramada o por demanda. El funcionamiento es sencillo: el master simplemente envía una réplica de todos los datos hacia los esclavos, en vez de solamente enviar los datos que fueron alterados desde el último Snapshot realizado. Se realiza un monitoreo de los cambios a los datos que son realizados en el master (inserciones, borrados y modificaciones de transacciones).

    Snapshot es útil cuando:

- Los datos son principalmente estáticos y no cambian a menudo. Cuando cambian, tiene más sentido publicar una nueva copia a los suscriptores.
- Es aceptable tener copias de datos que están fuera de fecha para un período de tiempo.
- Replicando pequeños volúmenes de datos en la que todo un refrescamiento de datos es razonable.
- Para decidir si la replicación Snapshot es apropiada, se debe considerar el tamaño de todo el conjunto de datos y la frecuencia de cambio en los datos. (Yissell Fernández Aguiar, 2008).

- **Administración-funcionamiento de la sobrecarga** (tecnología: ***disparadores/reglas bases de la replicación***):

  La función del disparador es ejecutar una acción cuando ocurre un evento en la base de datos, los eventos pueden ser de inserción, actualización o eliminación. Conjuntamente con las instantáneas, los disparadores son mecanismos asincrónicos que proporciona la base de datos como una manera de replicación de datos. Para mayor comodidad serán llamados triggers, ya que es el término por el cual son conocidos.

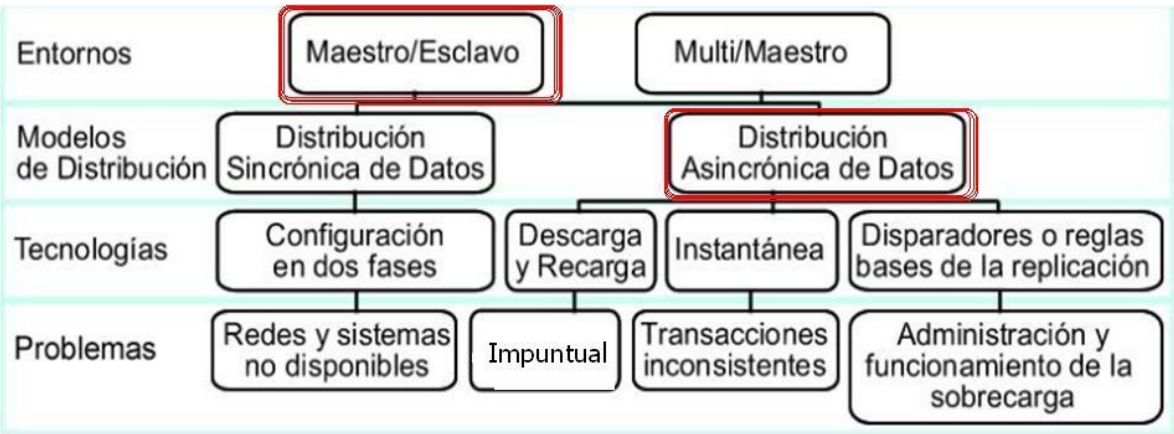

Fig. 2.5

Existen varios factores que vienen dados por los tipos de **Conflictos de Replicación** a tener en cuenta. Los conflictos pueden ocurrir cuando estamos trabajando en un ambiente de replicación que permite actualizaciones concurrentes sobre los mismos datos en las bases de datos.

- **Conflicto de actualización**
- **Conflicto de unicidad**
- **Conflicto de supresión**
- **Conflicto de orden**

Cada uno de estos conflictos vienen dados por:

- **Conflicto de actualización**

Un conflicto de actualización ocurre cuando se produce la replicación de una actualización (update) sobre un registro con otra actualización (update) sobre el mismo registro. Este conflicto ocurre cuando dos transacciones originadas desde distintos sitios actualizan el mismo registro, en forma cercana en el tiempo.

Posible solución:

Prioridad: Cada servidor obtiene una prioridad única, y el servidor de mayor prioridad “gana”, respecto a aquellos con prioridad menor.

Timestamp: La más nueva o la más antigua de las modificaciones es la considerada correcta, y por defecto, si no se eligió ninguno de los criterios, “gana” la más nueva.

Particionamiento de datos: Se garantiza que cada registro sea manipulado por un único servidor, lo que simplifica la arquitectura.

- **Conflicto de unicidad**

El conflicto de unicidad sucede cuando la replicación de un registro intenta violar una restricción de integridad, ya sea por llave primaria o única (Primary Key o Unique). Por ejemplo considere lo que sucede, cuando dos transacciones originadas de dos sitios diferentes, cada una inserta un registro, en su respectiva tabla replicada, con el mismo valor de clave primaria. En ese caso ocurre un conflicto de unicidad.

Posible solución:

Para cada servidor brindar un rango distinto de números para los generadores de clave (secuencias).

Agregar el identificador del servidor a la clave primaria.

Replicar en tablas separadas, y acceder a los datos a través de una vista formada por la unión de ellas. Para resolver el conflicto de potenciales claves duplicadas en la unión se usará una pseudo columna que representa la BD fuente.

- **Conflicto de supresión**

Un conflicto de supresión ocurre cuando dos transacciones originadas de sitios diferentes, una de ellas intenta borrar un registro, y la otra actualizar o borrar el mismo registro, ya que en este caso el registro no existe, tanto para ser actualizado como borrado.

Posible solución:

Para evitar este tipo de conflictos, una posible solución es que los sitios marquen lógicamente los registros a ser borrados y que periódicamente el sitio maestro corra un proceso que realice el borrado (“delete”) físico de los datos, es decir desde los sitios replicados no se puede ejecutar una sentencia para hacer el borrado de los datos (delete).

- **Conflicto de orden**

Los conflictos de orden pueden ocurrir en ambientes de replicación con tres o más sitios maestros. Si la propagación al sitio maestro X, está bloqueada por alguna razón, entonces la replicación de modificaciones en datos puede seguir siendo propagada a través de los otros sitios maestros; al finalizar la propagación estas modificaciones debieron ser propagadas al sitio X en un orden diferente a como ocurrieron en los otros sitios maestros, pudiendo producirse un conflicto.

Posible solución:

Este tipo de conflicto suele resolverse asignándole distintas prioridades a los sitios maestros, ordenando las transacciones de acuerdo a ésta.

2.6 Mecanismos de Replicación de Datos

Mecanismo, puede definirse como: la manera de producirse o de realizar una actividad. Entre los mecanismos existentes, se encuentran los de replicación de datos, tales como: **Replicate de CapeSoft** (replicación de del Software CapeSoft), este permite sincronizar o replicar datos de manera automática e independientemente del driver de base de datos que se esté utilizando. Es un mecanismo de replicación de datos entre dos o más bases de datos. Genera un registro de las adiciones, bajas y cambios realizados en la base de datos, para posteriormente, utilizando un administrador de transporte, exporta todos los cambios realizados a la base de datos a otro sitio,

donde las adiciones, bajas y cambios son importados a la base de datos elegida. Todo esto es realizado de manera automática sin que los usuarios tengan que realizar acción alguna adicional.

Además permite configurar el Proceso de los archivos de bitácora (importar / exporta) tan seguido como se requiera.

Principales Características:

- Rastreo de todas las Adiciones/Bajas y Cambios
- Soporta Logout/Rollback/Commit
- Programación variable de la Sincronización
- Exportación completa de Datos
- Replicación selectiva de Tablas y/o Campos
- Replicación Completa Bi-direccional
- Control de Versión de Sincronización
- Compatible con: Clarion 4, 5 y 5.5/16 y 32 bits/ABC o Procedural/Web Builder/ClarioNet/DLL o Local/Multi DLL.

[12]

Otro mecanismo es el “**Mecanismo Modelo-Entorno-Replicación (MMER)”,** el cual surge a raíz de la experiencia del estudio de otros mecanismos y su funcionamiento como realización de actividades y consiste en un proceso del cual se realiza una replicación paso a paso según las espectativas y lo que desee el suscriptor. Influye grandemente lo que el publicador permita replicar, puesto que este mecanismo no es de tipo replicación completa o de tipo backup. Puede realizarse a nivel de modelo síncrono y también de acuerdo al entorno de réplica definido, para conformar el algoritmo a utilizar para este proceso, en la replicación de datos.

Pasos (Algoritmo):

- Entendimiento y Confirmación.

- Selección del modelo de replicación.
- Selección del entorno de réplica.
- Tecnología de replicación a desarrollar.
- * Selección de posibles procesos externos. *
- Selección de la herramienta de replicación a utilizar.
- Realización del proceso de replicación de datos.

* *Opcional

2.7 Fases generales para implementar y supervisar la replicación

A pesar de que existen varias formas de implementar y supervisar la replicación, y el proceso de replicación es diferente según el tipo y las opciones elegidas, en general, la replicación se compone de las siguientes fases:

- Configuración de la replicación.
- Generación y aplicación de la instantánea inicial.
- Modificación de los datos replicados.
- Sincronización y propagación de los datos.

Configuración de la replicación:

La replicación de datos puede ser configurable, de acuerdo a los intereses del cliente y a qué se enfocaría principalmente su réplica.

2.8 Consideraciones finales en cuanto a replicación

La replicación es muy útil para mejorar la disponibilidad de datos, lo cual pudiera llevarse al caso extremo, conocido como bases de datos distribuidas replicadas totalmente, el cual consiste en la replicación de la base de datos completa en cada sitio en el sistema distribuido y garantiza notablemente la disponibilidad de datos, pues el sistema puede continuar operando cuando exista en servicio al menos uno de los servidores. En otros casos, como es el de esta investigación, la

replicación se efectúa para obtener datos principalmente de una fuente de almacén. Desde aquí, del servidor maestro, se propagan los datos necesarios y sólo estos a los esclavos para ser utilizado por este, y no como fuente de respaldo para el maestro.

En este caso, la réplica de datos, contenida en los esclavos, no les es necesaria ni imprescindible a los maestros. La información contenida en los esclavos, luego de efectuarse la réplica, le es de utilidad solo al cliente.

2.8.1 Ambientes de Replicación.

Un ambiente de replicación es una configuración de dos o más sitios mediante un escenario peer-to-peer. Cada sitio es un par que contiene un motor de replicación y una BD simple o compartida. El motor de replicación puede residir en la misma computadora que la BD asociada o en una computadora separada. En cada caso, la BD debe ser accesible por el cliente del motor de replicación (ej. ODBC1).

Cada sitio almacena solo los datos que requieren los usuarios locales. En segundo plano, el motor de replicación gestiona los cambios realizados a la BD sincronizando las actualizaciones de los datos con otros sitios activos en la red.

Una red de replicación puede ser de los siguientes tipos:

- ✓ Homogénea: Se replican datos entre BD con gestores y plataformas del mismo tipo (ej. PostgreSQL + Linux ←→ PostgreSQL + Linux).
- ✓ Homogénea con diferentes plataformas: Se replican datos entre BD con gestores del mismo tipo y plataformas diferentes (ej. PostgreSQL + Windows ←→ PostgreSQL + Linux).

Una red de replicación requiere una estructura básica TCP/IP que posibilite una comunicación efectiva y eficiente entre todos los sitios. La red de replicación por sí misma constituye estructuralmente una red virtual que se coloca por encima de la red física.

2.9 Minería de Datos

La minería de datos es un proceso opcional que puede facilitar y contribuir al desarrollo de esta investigación, puesto que no todos los datos que se obtendrán del servidor maestro son necesarios para los esclavos. No se precisa ser utilizado como un backup, ni un respaldo para la BD origen. Sólo se desean replicar aquellas informaciones que son de utilidad para el cliente. Este proceso se lleva a cabo mediante la minería de datos (DM, Data Mining) que consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos, y/o los datos relevantes. Es una poderosa tecnología nueva con gran potencial para ayudar a las compañías a concentrarse en la información más importante de sus Bases de Información (Data Warehouse).
Para un experto, o para el responsable de un sistema, normalmente no son los datos en sí lo más relevante, sino la información que se encierra en sus relaciones, fluctuaciones y dependencias.
Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las base de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las base de datos.

2.9.1 Pasos Generales

Un proceso típico de minería de datos consta de los siguientes pasos generales:

1. Selección del conjunto de datos, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo, como posiblemente al muestreo de los registros disponibles.
2. Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).
3. Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.
4. Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.

[13]

Evaluar los resultados contrastándolos con un conjunto de datos previamente reservado para validar la generalidad del modelo.

Si el modelo final no superara la evaluación, el proceso se podría repetir desde el principio o, si el experto lo considera oportuno, a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido.

2.9.2 Fases en la Minería de Datos

Se le denomina fase al asunto o paso dentro del proceso. Un proyecto de minería de datos tiene varias fases necesarias que son, esencialmente:

- Comprensión del negocio y del problema que se quiere resolver.
- Determinación, obtención y limpieza de los datos necesarios.
- Creación de modelos matemáticos.
- Validación, comunicación, etc. de los resultados obtenidos.
- Integración, si procede, de los resultados en un sistema transaccional o similar.

[14]

La relación entre todas estas fases sólo es lineal sobre el papel. En realidad, es mucho más compleja y esconde toda una jerarquía de sub-fases. A través de la experiencia acumulada en

proyectos de minería de datos se han ido desarrollando metodologías que permiten gestionar esta complejidad de una manera más o menos uniforme. Ejemplos de ellas son CRISP-DM y SEMMA.
Ver estas fases abordadas en el capítulo 3.

2.9.3 Técnicas de minería de datos

Las técnicas de la minería de datos provienen de la Inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados.

Las técnicas más representativas son:

- Redes neuronales.- Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida. Algunos ejemplos de red neuronal son:
  - El Perceptrón.
  - El Perceptrón multicapa.
  - Los Mapas Auto-organizados, también conocidos como redes de Kohonen.
- Árboles de decisión.- Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos:
  - Algoritmo ID3.
  - Algoritmo C4.5.
- Modelos estadísticos.- Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

- Agrupamiento o Clustering.- Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:
  - Algoritmo K-means.
  - Algoritmo K-medoids.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados (Weiss y Indurkhya, 1998):

- Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.
- Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.
  [14]

2.9.4 Minería de datos y otras disciplinas análogas

De la informática:

De la informática toma las siguientes técnicas:

- **Algoritmos genéticos**: Son métodos numéricos de optimización, en los que aquella variable o variables que se pretenden optimizar junto con las variables de estudio constituyen un segmento de información. Aquellas configuraciones de las variables de análisis que obtengan mejores valores para la variable de respuesta, corresponderán a segmentos con mayor capacidad reproductiva. A través de la reproducción, los mejores segmentos perduran y su proporción crece de generación en generación. Se puede además introducir elementos aleatorios para la modificación de las variables (mutaciones). Al cabo de cierto número de

iteraciones, la población estará constituida por buenas soluciones al problema de optimización, pues las malas soluciones han ido descartándose, iteración tras iteración.

- **Inteligencia Artificial**: Mediante un sistema informático que simula un sistema inteligente, se procede al análisis de los datos disponibles. Entre los sistemas de Inteligencia Artificial se encuadrarían los Sistemas Expertos y las Redes Neuronales.

- **Sistemas Expertos**: Son sistemas que han sido creados a partir de reglas prácticas extraídas del conocimiento de expertos. Principalmente a base de inferencias o de causa-efecto.

- **Sistemas Inteligentes**: Son similares a los sistemas expertos, pero con mayor ventaja ante nuevas situaciones desconocidas para el experto.

- **Redes neuronales**: Genéricamente, son métodos de proceso numérico en paralelo, en el que las variables interactúan mediante transformaciones lineales o no lineales, hasta obtener unas salidas. Estas salidas se contrastan con los que tenían que haber salido, basándose en unos datos de prueba, dando lugar a un proceso de retroalimentación mediante el cual la red se reconfigura, hasta obtener un modelo adecuado.

[15]

2.9.5 Tendencias

La Minería de Datos ha sufrido transformaciones en los últimos años de acuerdo con cambios tecnológicos, de estrategias de marketing, la extensión de los modelos de compra en línea, etc. Los más importantes de ellos son:

- La importancia que han cobrado los datos no estructurados (texto, páginas de Internet, etc.)

- La necesidad de integrar los algoritmos y resultados obtenidos en sistemas operacionales, portales de Internet, etc.
- La exigencia de que los procesos funcionen prácticamente en línea.
- Los tiempos de respuesta. El gran volumen de datos que hay que procesar en muchos casos para obtener un modelo válido es un inconveniente; esto implica grandes cantidades de tiempo de proceso y hay problemas que requieren una respuesta en tiempo real.

[14]

2.10 Minería de datos espaciales (MDE)

La minería de datos espaciales (MDE) es el descubrimiento de conocimiento implícito y previamente desconocido en base de datos espaciales (BDE). La MDE se refiere a la extracción del conocimiento, de las relaciones espaciales, o de otros patrones interesantes almacenados no explícitamente en BDE. Exige una integración de los datos que minan con tecnologías espaciales (Cabena et al. 1998). Puede ser utilizada para entender datos, descubriendo relaciones espaciales y no espaciales, construyendo bases de conocimiento espaciales, reorganizando preguntas y optimizando las bases de datos espaciales (Pineda et al. 1998).

Se espera que tenga usos amplios en SIG, geo-marketing, detección remota, exploración de imágenes en bases de datos, proyección de imágenes médicas, navegación, control de tráfico, estudios ambientales, y muchas otras áreas donde se utilizan los datos espaciales.
El conocimiento a ser descubierto en los datos espaciales puede ser de varios tipos, como características representativas, estructuras o agrupamientos, asociaciones espaciales, solamente por mencionar algunos.

2.10.1 Métodos de MDE

Los métodos de MDE son aplicados para extraer conocimiento interesante y regular. Estos métodos pueden ser usados para entender los datos espaciales, descubrir relaciones entre datos espaciales y no espaciales, reorganizar los datos en bases de datos espaciales y determinar sus características generales de manera simple y concisa (Michalski et al. 1998).

Existen cinco grupos de métodos de MDE:

- **Métodos basados en generalización**. Los cuales requieren la implementación de jerarquías de conceptos, en el caso de las bases de datos espaciales estas jerarquías pueden ser temáticas o espaciales. Una jerarquía temática puede ser ejemplificada al generalizar mango y piña a frutas. Una jerarquía espacial puede ser ejemplificada generalizando varios puntos en un mapa como una región y un grupo de regiones como un país.

- **Métodos de reconocimiento de patrones**. Estos pueden ser usados para realizar reconocimientos y categorizaciones automáticas de fotografías, imágenes y textos, entre otros.

- **Métodos que usan agrupamiento**. Consisten en crear agrupaciones o asociaciones de datos, cuando en estos existan nociones de similaridad (por ejemplo, distancia Euclidiana). Clustering es el proceso de agrupar datos en grupos o clusters de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y baja con objetos de otros clusters.

- **Métodos explorando asociaciones espaciales**. Permiten descubrir reglas de asociaciones espaciales, es decir, reglas que asocien uno o más objetos espaciales con otro u otros objetos espaciales (X -Y (c%)), donde X y Y son un conjunto de predicados espaciales o no espaciales y c% es la confianza de la regla. Su aplicación está en bases de datos grandes, donde puede existir una gran cantidad de asociaciones entre los objetos, pero la mayoría de ellos serán aplicables solamente a un pequeño número de objetos, teniendo en cuenta que la confianza de la regla puede ser baja.

- **Métodos que utilizan aproximación y agregación.** Descubren conocimiento en base a las características representativas del conjunto de datos. La proximidad agregada es la medida de proximidad del sistema de puntos en el grupo en base a una característica en comparación con el límite del grupo y el límite de una característica. Las consultas de proximidad solicitan objetos que se hallen cerca de una posición específica

[16]

En Fig. 2.6 se observa la clasificación de los métodos de minería de datos espaciales:

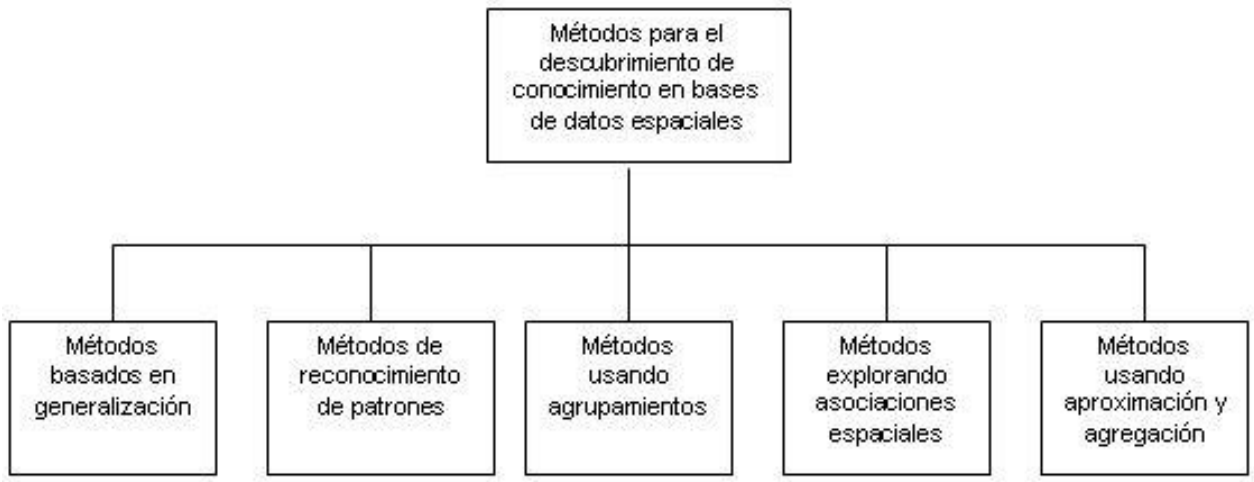

Fig. 2.6. Métodos para el descubrimiento de conocimiento en BDE. Tomado de J. Adhikary, 2001

2.11 Conclusiones

En el desarrollo de este capítulo se hizo un análisis sobre los modelos de replicación, donde se expusieron sus características principales. Además se determinaron los factores principales a tener en cuenta para elegir el modelo de datos a utilizar y se estudió el impacto de la minería de datos sobre la solución que se propone.

CAPÍTULO 3: “Propuesta y Documentación de la Solución”.

3.1 Introducción

En este capítulo se desarrolla la documentación y propuesta de la solución de la investigación. Se desarrolla un análisis del mecanismo de replicación y técnica de minería de datos seleccionados, así como a las fases de esta, desglosadas acorde a la situación de esta investigación. Mediante esta propuesta, quedaría conformada, lo que sería, la estrategia a seguir para replicar datos espaciales en Sistemas de Información Geográfica.

3.1 Análisis del mecanismo de replicación seleccionado

Una vez analizados los mecanismos y técnicas de minería de datos, es el momento propicio para definir cuáles de ellos utilizar, con el objetivo de trazar la estrategia de replicación de datos espaciales en SIG.

El mecanismo **Mecanismo Modelo-Entorno-Replicación (MMER)** consiste en una explicación paso a paso de cómo se produciría la replicación de datos relevantes para el cliente, según sus espectativas. Se realiza a nivel de modelo asíncrono. De acuerdo al entorno de réplica, el ideal a utilizar es el mastro-esclavo, por las características de sólo lectura que presenta el servidor receptor de la información. Este mecanismo consta de los siguientes pasos

3.1.1 Entendimiento y Confirmación

Realizar una replicación de una BD a otra, supone un proceso habitual y simple, pero enfocada al problema fundamental, el cual radica en la necesidad que hoy presenta el GSIG de encontrar un mecanismo factible para que de manera independiente a la entidad a la que se le esté personalizando una solución, se pueda acceder tanto al conjunto de datos espaciales como a los datos socioeconómicos que sean de mayor relevancia para el cliente, garantizando que no se

modifiquen las bases de datos originales y que la calidad del sistema no se vea afectada y sea mucho más exigente.

Mediante este mecanismo es posible establecer el proceso de replicación, puesto que además garantiza que no se violen aspectos fundamentales como no modificar los datos del servidor maestro.

3.1.2 Selección del modelo de replicación

Después de realizar el análisis correspondiente, descubrir y decidir sobre los modelos de replicación para el cual se llevará a cabo la replicación, se llega a la conclusión de que es por medio del **asíncrono**. Para ello se realizó un análisis de los diferentes factores que propiciaron a la toma de decisión sobre cuál de los modelos utilizar.

Una vez realizado este tipo de operación, sea para una o varias personas, la BD destino necesita alimentarse de esa nueva información socioeconómica. ¿De qué forma lo realizará? ¿Cómo? Puesto que pude darse el caso de que las BD no siempre estarán conectadas ni sincronizadas, lo que lleva a vincular los factores relacionados con el entorno de red, esta actualización de un nuevo registro se realiza de forma asíncrona, y puede ocasionar otro de los factores relacionados con los problemas de acuerdo con la tecnología de replicación utilizada. Estos problemas, pueden originarse en dependencia del manejo y uso de: la descarga/carga, instantánea y disparadores/reglas bases de la replicación. Una de las características del modelo de replicación asíncrono, y lo que lleva a definir concretamente su uso, es que puede efectuar la replicación de una manera que garantiza la efectividad en consecuencia de cómo los datos socioeconómicos son distribuidos.

El modelo seleccionado para efectuar la replicación es el asíncrono, porque la tecnología de replicación asíncrona es la más reciente alternativa para proporcionar tolerancia a fallos en servidores y almacenamiento en red. Implantando replicación asíncrona, se puede conseguir un acceso a los datos cercano al tiempo-real sin afectar a las aplicaciones.

¿Necesita el esclavo estar conectado todo el tiempo al maestro? ¿Por qué? Es en dependencia de lo que el cliente mismo establezca a la hora de realizar el proceso de replicación, y sobre lo que en es momento defina como prioridad. Si el servidor esclavo no fuera necesario todo el tiempo sincronizado al maestro, trae como beneficio de que el esclavo puede caer o quedar desconectado durante horas o días, luego reconectar y leer las actualizaciones. Por ejemplo, puede preparar una relación maestro/servidor mediante una red que sólo esté disponible casualmente y durante cortos periodos de tiempo.

3.1.3 Selección del entorno de réplica

Mediante el entorno, se puede conocer si es prudente o no escribir sobre ambos servidores, y sobre la direccionalidad. Definido el problema en cuestión, y ver todo lo relacionado sobre de donde y hacia dónde se realizará la replicación, se define que el entorno para efectuar la replicación es el **maestro-esclavo**, por las características asociadas en cuanto a escribir únicamente en el esclavo y por la unidireccionalidad del intercambio de información.

Dadas las características de cómo se llevaría a cabo la replicación para este tipo de problema en cuestión, se toma como entorno de replicación de datos al **Maestro-Esclavo**, por una serie de características que lo hace distinguir del otro entorno Maestro-Maestro. Dichas características cuentan sobre la base de que el cliente, es decir, la BD que desea alimentarse de la replicación, solamente tendrá acceso lectura a la BD de origen de datos, puesto que no se pueden modificar los datos almacenados realmente, porque afectarían un conjunto de aplicaciones que se alimentan de esa BD maestra. Otro de los factores que genera, es que el cliente no puede tener acceso de escritura a la de origen, es que alterando cualquier información, trae como consecuencia que se pierde la disponibilidad y la integridad de la información almacenada.

Esta replicación se llevaría a cabo mediante este modelo, que refleja los principios por los cuales esta investigación surge, garantizar los datos en la BD origen y que la calidad del sistema no se vea afectada.

Debido a que se realiza desde los maestros hacia los esclavos, la replicación se efectúa de forma unidireccional, que tiene beneficios para la robustez y velocidad del sistema. Al estar funcionando un servidor esclavo sin molestar al maestro, este continúa procesando actualizaciones u operaciones según sea la operación, mientras se realiza la replicación.

3.1.4 Tecnología de replicación a desarrollar

La tecnología, viene dada, explícitamente sobre cómo, cuándo y desde donde se realizará la replicación.

Las tecnologías asociadas a este tipo de replicación asíncrona serán las siguientes:

- Descarga/recarga.
- Instantánea.
- Disparadores/reglas bases de la replicación

El cliente puede utilizar la **Descarga/recarga** para actualizar los datos y establecer el proceso de replicación por medio de dispositivo de almacenamiento masivo, como es el caso de una Flash Memory, CD, DVD u otro medio de transporte de información. Estos archivos pueden ser de extensiones “.xml” o “.backup”. Trae como consecuencia que los datos se actualicen impuntualmente. Asumiendo que en un momento dado, no exista conexión o no se pueda establecer la sincronización mediante la red entre los servidores, se puede hacer un uso apropiado de esta tecnología.

En caso distinto, cuando es posible el establecimiento de la sincronización de las BD y la mayoría de los datos no cambian con frecuencia, el empleo de la tecnología **instantánea** es posible en este tipo de análisis, lo que produce que se repliquen pequeñas cantidades de datos; las BD con poca frecuencia están desconectadas, si se diese el caso, y es aceptable la configuración de un período de latencia largo (la cantidad de tiempo que transcurre entre la actualización de la información entre las BD).

El uso de la tecnología de **disparadores/reglas bases de la replicación**, es la ideal para cuando exista el caso de modificación constantes en el servidor maestro, en este tipo de replicación, de inserción, actualización o eliminación. La función del disparador es ejecutar una acción cuando

ocurre uno de esos eventos en la BD. Conjuntamente con las instantáneas y descarga/recarga, los disparadores son otras de las tecnologías en los mecanismos asincrónicos que proporciona la BD como una manera de replicación de datos para su funcionalidad en este problema de investigación.

Inicialmente la tecnología a utilizar será la Instantánea, puesto que los datos deben replicarse primeramente al esclavo, y luego para actualizar estos, se empleará la tecnología Disparadores.

3.1.5 Selección de posibles procesos externos

Un proceso externo para realizar la replicación, es la minería de datos, haciendo posible la obtención y selección de información que sólo le es relevante al cliente y permitiendo que no se realice un back-up de una BD a otra. Este proceso se llevaría a cabo sobre la búsqueda de conocimientos que estrían ocultos dentro de las BD.

3.1.5.1 Análisis de la técnica de minería de datos

Mediante la técnica de Agrupamiento o Clustering, de extracción de conocimiento, son aplicados los pasos a seguir para la realización de la minería de datos.

Este proceso de minería de datos pasa por las siguientes fases (Fig. 3.1):

3.1.5.2 Filtrado de datos

Luego de la comprensión del problema que se quiere resolver, es esperado que el formato de los datos contenidos en las base de datos casi nunca es el deseado o idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos "en bruto". Mediante el pre-procesado, se filtran los datos, de forma que se eliminan valores incorrectos, no válidos, desconocidos, según las necesidades y el algoritmo a usar. Se obtienen muestras de datos en busca de una mayor velocidad de respuesta del proceso, o se reducen el número de valores posibles, mediante redondeo y/o clustering.

¿Qué datos realmente se desean replicar? Mediante esta interrogante, es donde comienza el proceso de fase de filtrado de datos. La información a la que se accederá será la de mayor relevancia para el cliente. Esto implicaría filtrar todos los datos, de manera que si un cliente necesita algún dato que sea de carácter muy confidencial o afecte la seguridad de la base de datos socioeconómica, se le denegaría la petición. Se buscarían métodos alternativos o de carácter oficial, para informarle que este tipo de información ha sido denegada por no cumplir con lo establecido.

Para definir o saber qué información el cliente realmente puede replicar, según sus necesidades, se establecerían modelos de acceso a los datos. Una vez definido y verificado estos modelos, y realizada la **determinación, obtención y limpieza de los datos necesarios**, se procedería a la fase de selección de variables, para reducir el tamaño de los datos.

3.1.5.3 Selección de variables

Después de realizar la limpieza de los datos, en la mayoría de los casos se tiene una gran cantidad de variables o atributos. La selección de características reduce el tamaño de los datos, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería; seleccionando las variables más influyentes en el problema. Estas variables surgen o se definen en dependencia de la dimensión de la replicación a aplicar por cada cliente.

Para la selección de los atributos, se utilizan fundamentalmente dos métodos:

- Aquellos basados en la elección de los mejores atributos del problema:

Los principales atributos seleccionados, son los de mayor relevancia. Para el caso de los datos socioeconómicos, estas variables pueden ser las relacionadas con los identificadores, las llaves primarias que identifiquen fielmente a datos o conjuntos de estos, en conjunto que se desea replicar.

- Aquellos que buscan variables independientes mediante tests de sensibilidad, algoritmos de distancia o heurísticos.

Al seleccionar las variables más influyentes, mediante **la creación de modelos matemáticos** se procede a la esencia de la Minería de Datos, consistiendo en la extracción del conocimiento.

3.1.5.4 Extracción de Conocimiento

La extracción del conocimiento, mediante el empleo de la técnica de minería de datos, es posible obtener un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Los modelos que se generan son expresados de diversas formas:

- Reglas
- Árboles
- Redes neuronales

También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre-procesado diferente de los datos.

3.1.5.5 Interpretación y Evaluación

Una vez obtenido el modelo, se procede a su validación; donde se comprueba que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos para buscar el que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Esta fase es la más importante, y es catalogada como de integración, si procede, de los resultados en un sistema transaccional o similar.

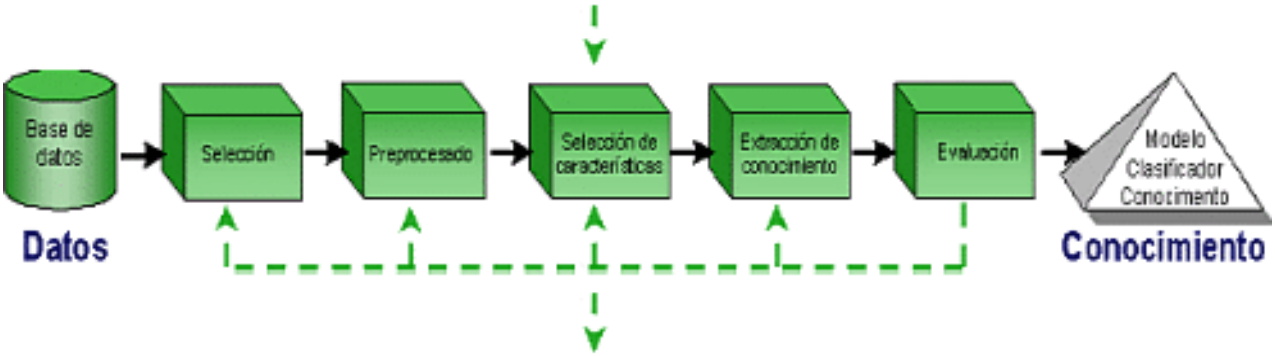

Fig. 3.1

3.1.6 Selección de la herramienta de replicación a utilizar

Para resolver el problema de cómo replicar y sincronizar los datos, es necesario el uso de una herramienta que permita esta operación, llegando a cumplir una gran función durante todo el proceso de replicación de datos. Mediante una herramienta que surte de intermedia entre los servidores y por razón de script, trigger y comandos, se hace posible el proceso de replicación.

A partir de los resultados que arrojaron las entrevistas a diferentes miembros de los proyectos de la UCI, que trabajan con replica de datos y del estudio de la propia replicación, se descubrió para beneficio de la investigación de una herramienta que por sus características y su maniobrabilidad, se puede llevar a cabo este desarrollo, y es: Magic@ Data Replication eXtensible Solution.

3.1.6.1 Magic@ Data Replication eXtensible Solution

Esta herramienta es un software no-propietario, bajo el lenguaje C#, desarrollada por el Ing. Jorge Landrian García, utilizada fundamentalmente en los proyectos de UCI, tales como “Registros y Notaria”, “ERP”, “Notaria” y también en parte de “Identidad”, permite la conexión al gestor de BD PostgreSQL, Oracle y SQL Server para extraer tanto las tablas de las cuales va a replicar datos hacia

otro servidor, como las tablas de otro servidor hacia las que va a mover datos, ofreciendo la replicación para el entorno maestro-esclavo. Incluso no sólo replicar algunos campos, sino poder combinar campos o cualquier expresión que se quiera replicar como una columna en la BD destino. Esta aplicación puede correr en cualquier nodo que tenga acceso a los dos gestores entre los cuales está replicando, puede ser tanto en la plataforma Windows donde se requiere el .NET Framework 2.0, o en la plataforma Linux con el Framework de mono (http://www.mono-project.com).

Debido a una real presencia de fallos cuando se realice el proceso de replicación, la herramienta, presta énfasis en:

- **Fiabilidad**, posibilitando que el proceso funcione sin interrupciones.
- **Seguridad**, haciendo posible que no se produzcan averías catastróficas.
- **Disponibilidad**, llegando a estar las BD disponibles el máximo de tiempo posible.
- **Mantenimiento**, fácilmente reparable.
- **Confidenciabilidad**, impidiendo el acceso no autorizado.
- **Integridad**, impidiendo la alteración inadecuada de la información.

La herramienta cumple con lo siguiente:

- Protección contra la pérdida de datos
- Recuperación de datos en tiempo real
- Acceso interrumpido en caso de falla

Funciona sobre los ambientes de replicación: homogénea:

- Homogénea con diferentes plataformas: Se replican datos entre BD con gestores del mismo tipo y plataformas diferentes (ej. PostgreSQL + Windows ←→ PostgreSQL + Linux).

3.1.6.2 Respuesta a Conflictos de Replicación

Este software ofrece una cierta respuesta antes los distintos tipos de conflictos de replicación, llegando a ser:

- Conflicto de actualización (update): si dos o más servidores distintos actualizan un registro en una sola tabla destinataria en forma cercana al tiempo, provocando un conflicto, esta herramienta asume que la más nueva de las modificaciones es considerada la correcta, y por defecto.
- Conflicto de unicidad (Insert, violación de Primary Key): consiste fundamentalmente sobre la base de cuando la replicación de un registro intenta violar una restricción de integridad, ya sea por llave primaria o única (Primary Key o Unique) provocado por las transacciones originadas de dos sitios diferentes, cada una debe insertar un registro, en su respectiva tabla replicada, con el mismo valor de clave primaria, este software intenta ofrecer una solución a este conflicto buscando si la llave primaria ya existente sea porque estuvo ahí presente antes de ocurrir el conflicto o mediante la introducción cercana al tiempo por otro servidor, entonces se produce la actualización de este registro.
- Conflicto de supresión (delete): ocurre cuando dos transacciones originadas de sitios diferentes, una de ellas intenta borrar un registro, y la otra actualizar o borrar el mismo registro, lo que provoca que la herramienta actúe de manera automatizada, adicionando o actualizando el valor que notifica sobre modificaciones del servidor en la tabla de control.

3.1.7 Realización del proceso de replicación de datos.

Descripción con la herramienta *Magic@ Data Replication eXtensible Solution.*

El proceso de replicación de datos se efectúa sobre la base de haber constituido cuál modelo (asíncrono) a desarrollar y el entorno (maestro/esclavo) sobre el cual estará jugando un papel primordial esta recolección de información. Para consumar los primeros movimientos, es necesario establecer o definir el ambiente (heterogéneo) en la compatibilidad de la herramienta de sincronización de los servidores. El intercambio de información se realiza de forma unidireccional,

desde la BD socioeconómica, que se encuentra ubicada en un gestor de BD: PostgreSQL, hacia la BD destinatario que posee GSIG.

La descripción de este proceso, es basada en el “Manual de configuración” de “*Magic@ Data Replication eXtensible Solution*” [17]. En lo siguiente, se describe detalladamente los pasos para la configuración y despliegue de esta solución de réplica, específicamente con servidores PostgreSQL; aunque la mayor parte del contenido es aplicable al resto de los gestores soportados por esta solución:

Esta herramienta consta de varios ficheros alojados en dos carpetas distintas:

1. Tools
    a. Aquí se encuentra, entre otros, el fichero “**RDBTool.exe**”, responsable de permitir la conexión al gestor de BD PostgreSQL para extraer tanto las tablas de las cuales va a replicar datos hacia otro servidor, como las tablas de otro servidor hacia las que se va a mover datos .

Fig. 3.3

    b. Como segundo fichero en importancia se encuentra: “**ReplicationSchemaTool.exe**” Esta herramienta permite diseñar gráficamente un esquema de replicación entre dos servidores, posibilitando definir el mapeo unidireccional de datos entre estos, además de las reglas que debe cumplir la información a replicar. Mediante este esquema es posible definir las tablas de donde se van a sacar los datos, hacia qué tablas van dirigidos estos, y las transformaciones necesarias para acoplarlos a las columnas de las tablas destino.

Fig. 3.4

2. DataSynchronizer

a. El fichero “**DataSynchronizer.exe**” es el principal dentro del proceso de replicación, haciendo posible toda la sincronización y ejecución de los comandos. Además, es una aplicación de consola que al ejecutarse según la configuración asignada ejecuta el proceso de sincronización entre los servidores.

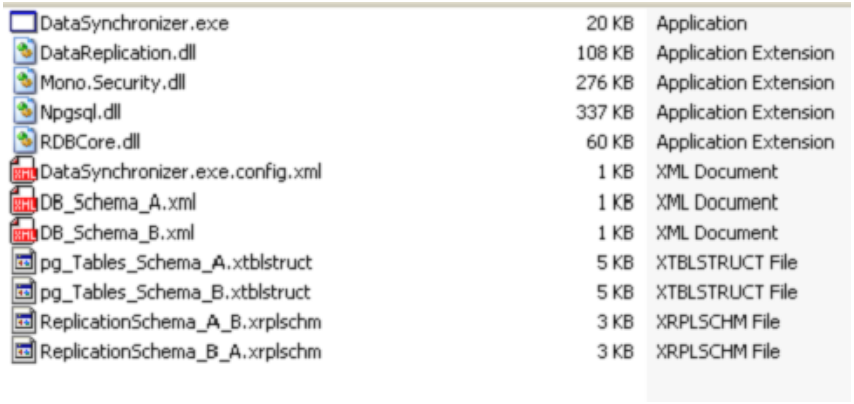

Fig.3.5

b. “**DataSynchronizer.exe.config.xml**” es un fichero de configuración. Este utiliza el concepto de base de datos local y base de datos remota, de donde se parte para entender la definición de subir información como el proceso de replicar desde la base de datos local hacia la remota, y bajar información como el movimiento de información del servidor remoto hacia el local.

3.1.7.1 Conexión entre las Bases de Datos.

Inicialmente las BD están desconectadas de ambos lugares, lo que conlleva a que como requisito indispensable en la interacción con la tecnología de **Triggers**, las BD deben estar conectadas, proceso por el cual es posible realizar con esta herramienta, para luego generar los respectivos scripts:

- Se procede a la conexión de los servidores, ejecutando el fichero “**RDBTool.exe**”: consiste en seleccionar las BD maestra y esclava, donde se especificarían los datos de los servidores a sincronizar. Utilizando la opción **Database -> Connection**, donde se mostrará una nueva ventana de configuración como en la Fig. 3.6

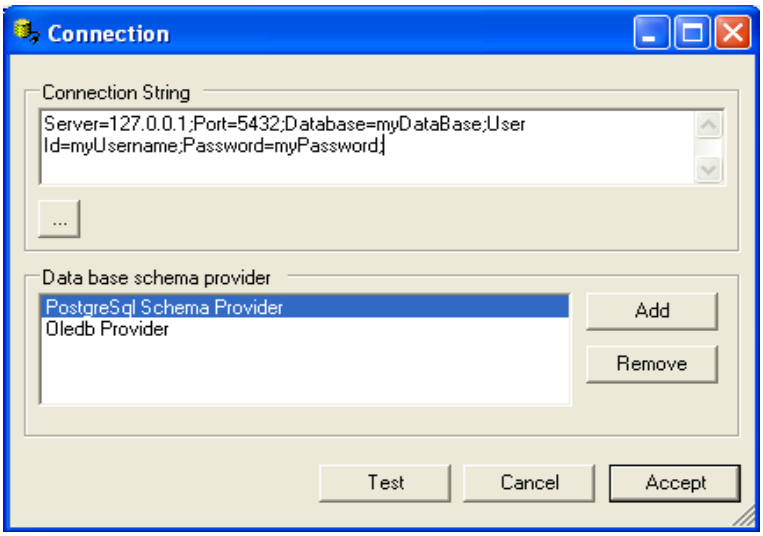

Fig 3.6

3.1.7.2 Selección de tablas.

Las tablas en PostgreSQL, están dentro de Esquemas, y estas a su vez dentro de las BD en los servidores, por lo cual, luego de realizarse un correcto proceso de conexión entre las BD, se llega al consenso de seleccionar las tablas que jugarán el papel primordial durante la replicación:

- El siguiente paso a realizar sería la adicción de tablas al proyecto utilizando la opción **Database ->Add Tables**. De acuerdo a las BD que se hayan conectado, aparecerán los esquemas de estas, y luego seleccionando estas, saldrá una distribución de tablas asociadas a este esquema. Las tablas aparecían en la herramienta, como se ilustra en la Fig. 3.7

  Quedarían distribuidas como se muestran a continuación:

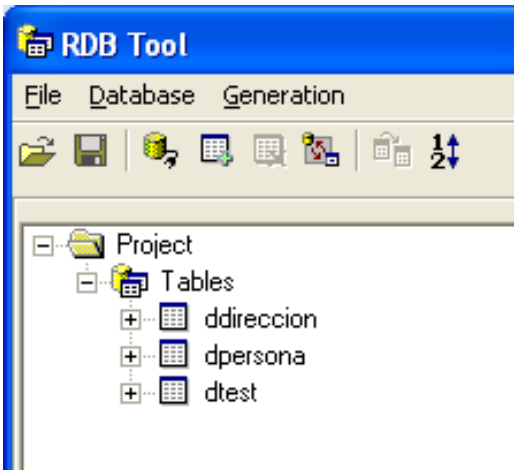

Fig. 3.7

Obteniendo estas tablas y comprobando que son las que intervendrán en la replicación, tanto de las de origen y destino, se procede a guardar estas en disco, utilizando la opción **File->Save** seleccione la extensión (Table structure **XML** file).

La creación y confección de este fichero es muy importante, pues será utilizado por “**ReplicationSchemaTool.exe**” para el mapeo y selección correcta de la relación de los campos a replicar.

3.1.7.3 Generación de scripts de configuración

Los scripts constituyen ficheros (de extensión **.sql**) que contienen consultas en lenguaje SQL y PgSQL, permitiendo desarrollar y crear nuevos elementos en las BD tales como: tablas, funciones, Triggers. La herramienta “Magic@ Data Replication eXtensible Solution.”, le da una correcta estructura y solución a la generación de estos ficheros, permitiendo de forma automatizada disponer de ellos para comenzar los primeros pasos importantes dentro de la replicación. Es muy importante antes de generar estos script de inicialización, configurar algunas propiedades que son necesarias. Para ello se utiliza la opción **Generation->Tables- >Replica Control->PostgreSQL->Custom Properties**.

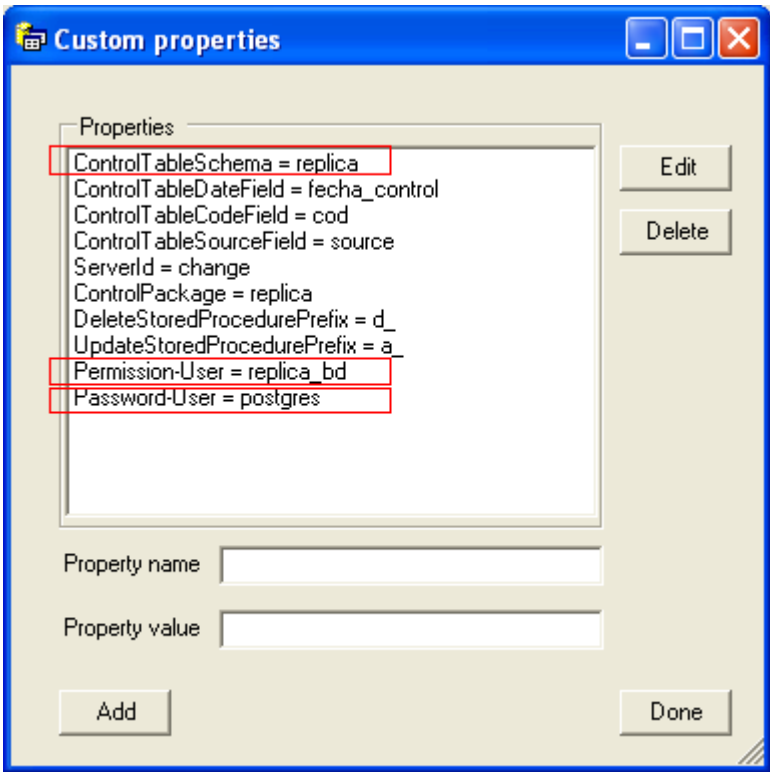

Fig. 3.8

Donde se exponen las siguientes propiedades:

- **ControlTableSchema**: se le asigna el nombre del esquema donde estarán contenidas las tablas de control de la réplica al igual que los triggers y funciones necesarias para su funcionamiento.
- **ControlTableDateField**: nombre de la columna, que recopila la fecha, de la tabla que controla la replicación.
- **ControlTableCodeField**: nombre de la columna, que almacena el tipo de operación, de la tabla que controla la replicación.
- **ControlTableSourceField**: nombre de la columna, que almacena el origen de los datos, de la tabla que controla la replicación.
- **ServerId**: por cada servidor que se tenga en la red de replicación, se le asigna un ServerId, consistente en un numérico único (1, 2, 3…) por lo tanto cada vez que se genere el script de los triggers para un servidor hay que actualizar esta propiedad con el ServerId correspondiente.
- **DeleteStoredProcedurePrefix:** prefijo para generar nombre de funciones que se encargan de eliminar una tupla.
- **UpdateStoredProcedurePrefix:** prefijo para generar nombre de funciones que se encargan de actualizar una tupla.
- **Permission-User**: se le asignan el usuario con el que se va a conectar el servicio de sincronización.
- **Password-User**: se le asignan la contraseña con el que se va a conectar el servicio de sincronización.
- **ControlPackage**: en el caso de esta propiedad es recomendable dejarla con este mismo nombre que viene por defecto: “replica”.

Los scripts que se generen, son ejecutados en los servidores correspondientes. Para formarlos, y ser usadas bajo PostgreSQL, se accede a **Generation->Tables->Replica Control->PostgreSQL**, y se eligen entre los 7 scripts, comenzando por el **00** y terminando en el **06**. Estos son:

- **pg_00_Script de Inicialización**
- **pg_01_Tablas de control**
- **pg_02_Triggers de control**:
- **pg_03_Funciones de Aplicacion de Cambios**
- **pg_04_Permisos**
- **pg_05_Borrar datos sincronizados tablas de control**
- **pg_06_Borrar datos tablas de control**

✓ La primera vez que se configura el servidor debe generar el script de Inicialización (**pg_00_Script de Inicialización**), este script crea el esquema donde se van a encontrar las tablas de control, triggers y funciones pertenecientes a la réplica. En la tabla de control de los servidores, **control_servers**, es donde se mantiene actualizada la fecha de sincronización con el resto de los servidores con los cual se sincroniza, como muestran las Fig. 3.9 y 3.10

	server [PK] character varying(256)	fecha_sinc_local timestamp(6) without time zone	fecha_sinc_remoto timestamp(6) without time zone	estado numeric(2,0)
1	B	2009-02-27 22:02:52.578	2009-02-27 22:02:53.031	1

Fig. 3.9 Tabla **control_server** del servidor A.

	server [PK] character varying(256)	fecha_sinc_local timestamp(6) without time zone	fecha_sinc_remoto timestamp(6) without time zone	estado numeric(2,0)
1	A	2009-02-27 22:02:53.031	2009-02-27 22:02:52.578	1

Fig. 3.10 Tabla **control_server** del servidor B.

Es muy importante percibir que el campo **server** donde almacena valores (de character varying (256)), significa el identificador del servidor al cual estará conectado.

✓ El proceso de mantener actualizadas las tablas de control se realiza mediante Triggers sobre las tablas que se replican. Estos Triggers se generan con la opción **pg_02_Triggers de**

control. En este caso es muy importante tener en cuenta la propiedad ServerId, anteriormente vista.

- ✓ En el caso de las funciones de aplicación de cambios, se parte de las tablas en las cuales va a almacenarse información proveniente de otros servidores. Es importante realizar este análisis por cada servidor de la red de replicación, teniendo siempre bien claro cuáles son las tablas de las que sale información y cuáles son las tablas en las cuales se almacena información proveniente de otros servidores. Es perfectamente válido que de una misma tabla tanto salga como entre información. Según este tipo de análisis siempre vamos a tener dos ficheros de estructura de tablas (creados con la herramienta: RDBTool.exe), aunque en el caso probable de que sean las mismas tablas que salen las que entran, se puede tener un sólo fichero (.xtblstruct). Para generar el script de las funciones de aplicación, se utiliza la opción **pg_03_Funciones de Aplicación de Cambios**. Es importante tener en cuenta las propiedades DeleteStoredProcedurePrefix y UpdateStoredProcedurePrefix, anteriormente vistas.

- ✓ En el proceso de sincronización, el componente **DataSynchronizer.exe** se conecta al servidor utilizando el usuario configurado en Permission-User para asignarle los permisos en donde se genera el script de permisos **pg_04_Permisos**. Este usuario juega un papel determinante y los permisos son otorgados de acuerdo al entorno de replicación seleccionado. En esta solución se llego al consenso del maestro/esclavo, por tal motivo, el usuario que se creará en el servidor esclavo que permita la conexión con el maestro, sólo podrá tener acceso de lectura o que es lo mismo la condición **SELECT** del SQL, mientras que el otro, que permite la conexión viceversa a este, tendrá acceso total, puesto que debe insertar (**INSERT**), actualizar (**UPDATE**) y eliminar (**DELETE**) aquellos datos por el cual se ofrece la replicación. Se requiere verificar y modificar según resulte conveniente el script **pg_04_Permisos** antes de ejecutarlo en el servidor, puesto que conlleva a un nivel de seguridad considerable y razón por la cual se deriva esta investigación.

- ✓ A medida que pase el tiempo y en dependencia de la cantidad de operaciones en la BD, el tamaño de las tablas de control irá aumentando progresivamente, el proceso de sincronización no elimina las tuplas de las tablas de control que ya han sido replicadas, con el objetivo de no afectar el rendimiento en el momento de la sincronización. Para la limpieza de estas tablas la herramienta generados funciones **ctrl_borrar_tablas_control()** y **ctrl_borrar_tablas_ctrl_null().** Estas funciones se generan en dos opciones por separado porque es importante tener en cuenta sobre qué tablas se generan estos scripts. La primera función **ctrl_borrar_tablas_control()** se genera con la opción **pg_05_Borrar datos sincronizados tablas de control**. El objetivo de esta función es eliminar todas aquellas tuplas en las tablas de control en las que la fecha de control sea menor que la mínima fecha de sincronización con los servidores externos. Por tanto, el conjunto de tablas a tener en cuenta en el momento de generar este script es la unión de todas las tablas de las cuales se replica información hacia otros servidores.

- ✓ En el caso de la función **ctrl_borrar_tablas_ctrl_null()** el objetivo es eliminar todas aquellas tuplas de control de las tablas hacia las cuales se replican datos desde otros servidores y a la vez estas tablas no están en el conjunto de las tablas de las cuales se replican datos hacia otros servidores, es decir todo aquello que entre y no salga.

  Partiendo de:
  E – La unión de todas las tablas hacia las cuales se replica información de otros servidores.
  S – La unión de todas las tablas desde las cuales se replica información hacia otros servidores.

  Las tablas a tener en cuenta en la generación de la función **ctrl_borrar_tablas_ctrl_null()** serían las pertenecientes al conjunto **E – S**. La opción para generar el script con esta función es **pg_06_Borrar datos tablas de control**.

3.1.7.4 Mapeo y selección de tablas fuentes/destino.

La replicación de los datos se puede desarrollar de varias formas, tales como la utilización de filtrado de datos, reglas de replicación, o simplemente haciendo uso de una replicación total de tipo espejo. Como se ha visto dentro de la problemática planteada en esta investigación, la replicación no debe estar enfocada a que sea de tipo “espejo”, lo que se hace necesario el empleo de esta herramienta facilitando este proceso, mediante el mapeo en el esquema de replicación.

Un esquema de replicación es el concepto que permite definir el mapeo unidireccional de datos entre dos servidores, además de las reglas que debe cumplir la información a replicar. Es posible definir las tablas de donde se van a sacar los datos, hacia qué tablas van dirigidos estos datos, y las transformaciones necesarias para acoplarlos a las columnas de las tablas destino. Luego se accede a la creación de un mapeo, lo que trae consigo la opción de seleccionar la tabla fuente y la tabla destino entre las cuales se desea realizar el mapeo de datos.

Al seleccionar una tabla destino se muestran los campos de esta tabla en la columna “Destination column”, como se muestra en la Fig. 3.11. Según la selección de la tabla fuente, en el área de “source table columns” se muestran las columnas de la misma. Mientras se tiene el foco en una de las cajas de “Source expression” se puede hacer doble clic sobre una de las columnas de la tabla fuente generándose automáticamente el nombre de esta columna en el cajón de texto, agilizando la construcción de expresiones fuentes las cuales son los valores que se extraerán de la tabla fuente y se asignan a las columnas destino.

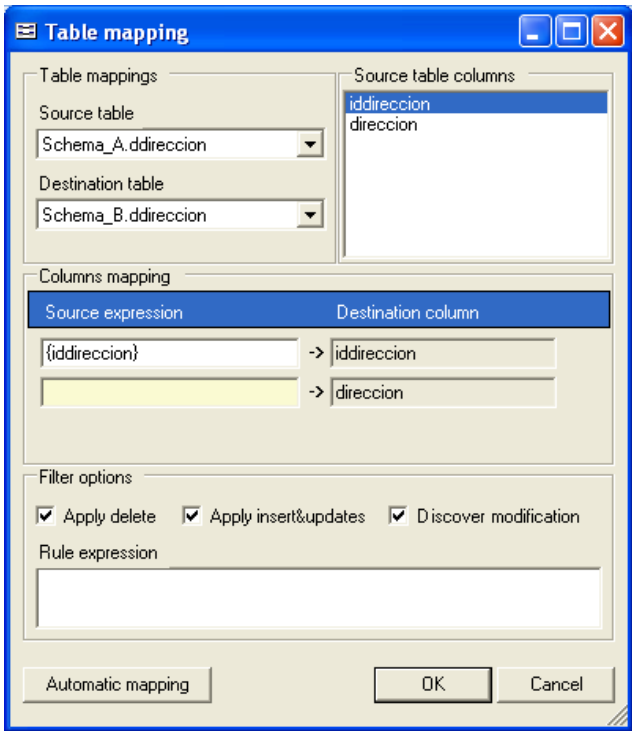

Fig. 3.11

Los campos que no se vayan a replicar, porque no le son significativos al cliente, serán dejados en blanco o no asociados mediante el mapeo. Esta representa una solución muy importante a la hora de determinar qué campos datos replicar y cuáles no.

3.1.7.5 Ejecución de los scripts de configuración

La ejecución de scripts, consiste en que ya contando con estos siete ficheros, una vez creados, son consumados en sus respectivos servidores. Los que son generados de la conexión por medio del servidor origen, deben ser ejecutados en este, lo mismo para los scripts del otro servidor en ese, significando contra que servidor se está sincronizando.

Por ejemplo, partiendo de que exista un servidor **A**, el cual brinda la información hacia un servidor **B** sobre ese se realizaron los pasos correspondientes para establecer la conexión con la BD, con la

correcta cadena de conexión, luego se seleccionan las tablas del **A**, logrando así poder contar con todos los scripts generados a partir de las tablas seleccionadas y de los del **00** al **06**. Se accede mediante una interfaz visual, ya sea el “PgAdmin III” o el “POSTGRESQL MANAGER 3”, al servidor **A**, luego se abren los scripts (o se cargan) y estos son ejecutados uno a uno en el orden en que fueron creados. Se ejecutan las consultas en ese orden y en todos los servidores que estarán implicados en el proceso de replicación

Como disimiles herramientas dentro de sus configuraciones y sus correctas ejecuciones, intervienen ficheros incluidos que proveen de informaciones necesarias a la hora de ser ejecutadas.
La herramienta que permite darle solución a esta problemática no se queda detrás y hace uso de los ficheros **.xml**, **.xrplschm**, **.xtblstruct** lo cuales deben ser reconfigurados con mucho cuidado y precisión, de acuerdo lo que sugieren las etiquetas xslt y xml, sobre los orígenes de diferentes archivos.

Como sugerencia en el desarrollo de las BD de GSIG, existen algunos archivos donde en los xslt de generación de scripts se podría cambiar su mecanismo, por ejemplo, este es muy cómodo en el caso de que las tablas se les nombrara dpersona, destudiante y cuando se cambiara la primera letra es muy fácil de identificar. En caso de que no sea así se podría cambiar como se va a generar este proceso, pero se debe hacer tanto en los xslt de generación de scripts como en los xmls de configuración.

Entre los ficheros .xml, se encuentra **DataSynchronizer.exe.config.xml** (como muestra la Fig.3.5), encargado de utilizar el concepto de base de datos local y base de datos remota, de donde se parte para entender la definición de subir información como el proceso de replicar desde la BD local hacia la remota, y bajar información como el movimiento de información del servidor remoto hacia el local.

3.1.7.6 Sincronización e inicio de la réplica

Cuando se sincroniza una réplica, al inicio se debe partir que las informaciones dos de las BD sean iguales, es decir, se parte de las BD conectadas y a partir de ahí es que se establece la replica, porque la razón de ser de esta es que todas las modificaciones hechas en una se apliquen en la otra.

Si este no fuese el caso, porque inicialmente se quiere copiar todos los datos de una BD en otra de una vez, existiendo estos en la de origen y no en la de destino:

- Se utiliza la réplica, de la forma en que se debe notificar los "id" de las tuplas a replicar en la tabla de control.
- Se insertan los "id" de la tabla de datos y se les ubica en el campo "source" de la tabla de control.
- El "id" de la BD en que están y en el campo "cod", se modifican a 0 (cero) significando que es cuando es un “insert” o “update”, de esta forma cuando se ejecute la réplica, la tabla detectará todo ello como modificaciones nuevas, la aplicará en la otra BD y los datos serán copiados

Como solución a ello, es posible empleando la siguiente consulta:

```
insert into esquema.ctabla_control (id_tabla, cod, source)
SELECT id_tabla, 0, 1
 FROM esquema.tabla
 where idtrabaj not in (SELECT id_tabla from esquema.ctabla_control)
```

La cual seria recomendable desarrollar una solución genérica, que incluya una opción, dentro de la herramienta, que permita generar esta consulta como unos script y ser utilizada en el servidor que publicará los datos.

Ya contando con que las BD tengan los datos necesarios y correspondientes, se procede a la replicación basada en la tecnología de Triggers, consistiendo en la ejecución del archivo

“**DataSynchronizer.exe**”. Cada vez que se quiera efectuar la replicación de los datos, se ejecuta ese fichero, informando de errores, si son creados, en el fichero: “**DataSynchronizer.exe.log**”

3.2 Conclusiones

Como conclusión al actual capítulo, se llega a la solución del problema en cuestión, lográndose la integración con la herramienta propicia para el establecimiento del proceso de replicación y la integración de los datos asociados necesarios para el cliente.

CONCLUSIONES GENERALES DEL TRABAJO

Una vez terminado el desarrollo de la investigación, es posible arribar a las siguientes conclusiones:

- Se realizó un análisis de las soluciones existentes en el campo de replicación de datos que sirvieron como punto de partida para el posterior desarrollo de la investigación.
- Durante la investigación se caracterizaron los diferentes modelos, mecanismos y entornos utilizados actualmente para la replicación de datos, por lo que se realizó la identificación y análisis de los más eficaces para llevar a cabo este proceso.
- Se determinó el empleo de un nuevo mecanismo de replicación propio a partir de las experiencias acumuladas de los utilizados actualmente, que satisficiera las necesidades de la investigación.
- Se utilizó la herramienta "*Magic@ Data Replication eXtensible Solution*", para llevar a cabo el proceso final de replicación bajo servidores PostgreSQL.
- Se logró solucionar el problema científico que dio origen a la investigación pues se cumplió el objetivo general trazado, al establecer una estrategia eficiente para las réplicas de datos espaciales en los Sistemas de Información Geográfica.

.

RECOMENDACIONES

Para extender el trabajo realizado, se recomienda la integración y desarrollo de la minería de datos, enfocadas a la búsqueda de conocimiento en cuanto al filtrado de datos automatizado. Además de la creación de una interfaz gráfica que permita al usuario interactuar entre la herramienta y los servidores, en cuanto a la replicación inicial y transporte de datos.

La solución puede extenderse a otros proyectos que necesiten hacer réplicas selectivas en cuanto a la información a ser replicada, es decir, no replicar todos los datos. Además puede presentarse en eventos que abarquen el tema de modo que se intercambie con expertos la solución, pudiendo por consiguiente profundizarse en la misma.

REFERENCIAS BIBLIOGRÁFICAS

1. *Naturaleza de los Datos espaciales*. [consultado 2009; Disponible en : http://www.udistrital.edu.co/comunidad/profesores/rfranco/datos_espaciales.htm.
2. *¿Qué es una Base de datos? - Definición de Base de datos*. [consultado 2009; Disponible en : http://www.masadelante.com/faqs/base-de-datos.
3. *Aspectos del Mecanismo de Replicación de Microsoft SQL Server 7.0.* [consultado 2009; Disponible en : http://gustavolarriera.tripod.com/doc/tech/sql7repl.pdf.
4. *Productos PostgreSQL – Cybercluster - Cybertec Schönig & Schönig GmbH*. [consultado 2009; Disponible en : http://www.postgresql-support.de/spanish/pr_cybercluster_s.html.
5. *Replicación de Datos : DBMoto*. [consultado 2009; Disponible en : http://www.hitsw.com/localized/spanish/products_services/dbmoto/dbmoto.html.
6. *Double-Take from Double-Take Software for Disaster Recovery...* [consultado 2009; Disponible en : http://es.doubletake.com/products/double-take/default.aspx.
7. *Replicación de SQL Server*. [consultado 2009; Disponible en : http://msdn.microsoft.com/es-es/library/ms151198.aspx.
8. *WhyFLOSS Conference 5º edición WhyFLOSS Conference 5º edición...* [consultado 2009; Disponible en : http://www.whyfloss.com/es/conference/corrientes08/getpdf/106.
9. *Deficicion de Modelo*. [consultado 2009; Disponible en : http://www.wordreference.com/definicion/modelo.
10. *Revista Uxi-numero 13*. [consultado 2009; Disponible en : http://www.scribd.com/doc/7006606/uxiVol203.
11. *FAQ/es - PostgreSQL Wiki*. [consultado 2009; Disponible en : http://wiki.postgresql.org/wiki/FAQ/es.
12. *Tribune Noviembre 2002 - Gopac Soluciones Integrales*. [consultado 2009; Disponible en : http://www.gopac.com.mx/tribune/noviembre02.htm.
13. *Business Intelligence Software - Nära Nivå | Business Intelligence ...* [consultado 2009; Disponible en : http://www.naraniva.com/business-intelligence-software/.
14. *Data Warehouse*. [consultado 2009; Disponible en : http://dev.url.edu.gt/moodle/file.php?file=/482/IntroduccionMD_enero_2009.ppt.
15. *Data Mining o Minería de Datos.* [consultado 2009; Disponible en : http://www.csae.map.es/csi/silice/DW2253.html.
16. *Mecanismos de solucion de controversias*. [consultado 2009; Disponible en : http://catarina.udlap.mx/u_dl_a/tales/documentos/msp/pech_p_ma/capitulo2.pdf.
17. Landrian García, J., *Manual de configuración de Magic@ Data Replication eXtensible Solution*. 2009.

BIBLIOGRAFÍA

[En línea] [Citado el: 23 de Enero de 2009.] http://www.geotecnologias.com/Documentos/GIS.pdf.

[En línea] [Citado el: 23 de Enero de 2009.] http://www.conida.gob.pe/capacitacion/sig.html.

[En línea] [Citado el: 22 de Enero de 2009.] http://wiki.postgresql.org/wiki/Preguntas_Frecuentes.

[En línea] [Citado el: 22 de Enero de 2009.] http://www.wordreference.com/definicion/modelo.

[En línea] [Citado el: 14 de Enero de 2009.] Adventures In Data Replication. *Network Computering.*

[En línea] [Citado el: 22 de Enero de 2009.] http://www.networkcomputing.com/603/603work1.html.

Elmasri, R y Navathe, S. 2000. *Fundamentals of database systems.Addison-Wesley.* Third Edition. 2000.

García, C. 1999. *Escenario de red para la supervisión de fallas en centrales telefónicas.* Universidad Central de "Marta Abreu" de Las Villa. Santa Clara : s.n., 1999. Informe Final de Tesis de Maestría en Computación Aplicada.

2008. Postgres-R. [Online] 2008. [Consultado: Enero 21, 2009.] http://www.postgres-r.org/documentation/terms_2.

WordReference.com. [En línea] [Citado el: 27 de Enero de 2009.] http://www.wordreference.com/definicion/sincr%C3%B3nico.

Cabena p., Hadjinian p., Stadler r., Verhees j. & Zanasi a. (1998) Discovering Data Mining. Prentice Hall.

Pineda l., Vega j. & Dorado a. (1998) Evaluación y Selección de una Técnica de Minería de Datos.. Facultad de Ingeniería Pontificia Universidad Javeriana.

Michalski r. S., Bratko i. & Kubat m. (1998) Machine Learning and Data Mining, John Wiley & Sons.

GLOSARIO DE TÉRMINOS

Transacciones: son grupos de operaciones que deben dar la apariencia de ser ejecutadas secuencialmente como una unidad. La definición de 'correcta' en una transacción se refiere al cumplimiento de las propiedades ACID.

Consistencia: "…Una transacción transforma un estado consistente de la base de datos en otro igual sin necesidad de conservar la consistencia en todos los puntos intermedios".

Durabilidad: Una vez que la transacción ha terminado, su efecto no puede perderse en caso de fallas del sistema; ni siquiera si la falla ocurre inmediatamente después de terminada la transacción.

Nodo: En términos de la propuesta para replicar base de datos fragmentadas, son dos bases de datos PostgreSQL que intervienen en la réplica.

Cluster: En términos de Slony-I, un cluster son dos bases de datos PostgreSQL que intervienen en la réplica.

Backup: copia de seguridad o copia de respaldo, se refiere a la copia de datos de tal forma que estas copias adicionales puedan restaurar un sistema después de una perdida de información.

ANEXOS

Entrevista

Para el desarrollo del presente trabajo de investigativo se realizó un estudio de las herramientas que intervienen en la replicación de datos, a través de varias entrevistas semi-estandarizadas realizadas a especialistas relacionados con este tema.

Las preguntas fueron las siguientes:

1. Se desea solucionar un problema que presenta el trabajo investigativo sobre replicación, principalmente en cómo hacer este tipo de proceso, para nada más replicar campos significativos de las tablas de una base de datos hacia otra. ¿la herramienta que ustedes utilizan en el proyecto permite este tipo de procedimiento?
2. ¿Existe alguna otra herramienta que ustedes conozcan que permita hacer este tipo de operación?
3. (Si son afirmativas ambas repuestas) ¿esta herramienta es libre?
   ¿Permitiría desarrollar la replicación en el entorno maestro/esclavo? Posibilitando la decisión del mecanismo para desarrollar la réplica de las BD.

Se efectuó la entrevista al proyecto Akademos, el cual arrojó el siguiente resultado de la entrevista y repuesta a las preguntas.

1. No, porque la herramienta que se emplea para solucionar la replicación de las bases de datos distribuidas, es el desarrollo de PyReplica la cual hace una replicación completa de datos.
2. No.

Extendiendo la entrevista a otros proyectos, se llego a contactar con el Diseñador de Bases de datos del proyecto: Sistema de Información de Pozos en Perforación.

1. No, porque la herramienta que en estos momento se encuentra utilizando el proyecto, figura el Slony-I, el cual efectúa una replicación de tipo espejo, lo que hace efectiva una replicación completa de toda la información, en la base de datos
2. No. Hasta ahora no conozco una herramienta que permita solucionar dicho problema.

Otro de los entrevistados, es el Ing. Jorge Landrian Garcia, crador y desarrollado de la herramienta "Magic@ Data Replication eXtensible Solution":

1. La solución que se utiliza actualmente en Registros y Notaría, en el ERP y también en parte de Identidad, que es la que yo desarrollé si permite hacer esto. Incluso no solo replicar algunos campos, sino poder combinar campos o cualquier expresión que se quiera replicar como una columna en la bd destino.
2. Sí, la herramienta "Magic@ Data Replication eXtensible Solution". Permite la replicación mediante el entorno maestro/esclavo sin afectar la seguridad del servidor maestro, además de que es un software no-propietario.