23

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
"ACATLAN"

LA LIMPIEZA DE DATOS, UN COMPONENTE DE LA ESTRATEGIA EMPRESARIAL PARA ASEGURAR LA CALIDAD DE INFORMACIÓN

TESINA

QUE PARA OBTENER EL TITULO DE
LICENCIADO EN MATEMATICAS APLICADAS Y COMPUTACIÓN

PRESENTA
IVOV JELEZOV DIMITAR

ASESOR: JUAN CARLOS RENDON

MAYO 2001

UNAM
CAMPUS ACATLAN

A mis padres Lilia e Ivo
Por darme la vida, por el largo camino que hemos recorrido juntos, por su amor, sacrificio, lucha, humanidad e infinito apoyo.
¡Eternamente gracias!

A mi hermana Rumiana
Por ser la amiga que siempre estará en mi corazón

A mis abuelos María y Dimitar, Nina y Dimitar.
Por ser un ejemplo, por su sabiduría, y porque siempre perdurarán en mi memoria y en la del tiempo como grandes personas

A mis tíos Kalin y Mariana
Por sus muestras de apoyo y cariño y porque somos una gran familia

A mi novia Rocío
Por su amistad, cariño, apoyo y amor que me ha brindado, durante mis estudios y por el futuro

A mi Asesor
Gracias Juan Carlos por el apoyo brindado, por el profesionalismo y por la confianza que tuviste hacia mí en el proceso de elaboración de este trabajo

A mis Sinodales
Gracias a Beatríz, Araceli, Georgina, Rubén y a todos los demás maestros y personal académico que me dieron la oportunidad de avanzar en mis estudios y por la enseñanza que me transmitieron.

A mis amigos
Gracias a Abigail, Angélica, Marco, Víctor, Rufino, Beto, Carlos, Rolando y Alvaro, por los inolvidables momentos que pasamos juntos

INDICE

INTRODUCCIÓN

La información siempre ha tenido un papel fundamental en el desarrollo de la ciencia, tanto en el pasado como en la actualidad. Hoy en día, la información se esta convirtiendo en un elemento fundamental del ambiente de los negocios. En un mundo de negocios que experimenta un desarrollo cada vez más dinámico la cantidad de datos que se captura, introduce, manipula y consulta dentro de los sistemas informáticos aumenta con cada día. Como consecuencia, la constante necesidad e una calidad creciente de los datos, se vuelve un aspecto mas importante a medida que crece el grado de complejidad de los recursos informáticos La calidad de los recursos informáticos en las diferentes áreas de una empresa es tan dispareja como diversos son los datos. Es un hecho que estas diferencias de nivel crecen a raíz de que aumenta la cantidad de nuevas técnicas y productos informáticos. Los análisis hechos muestran que la calidad de los datos en una empresa puede ser y casi siempre es mas baja y más dispareja de lo que se supone. Algunas otras investigaciones sobre los sistemas informáticos muestran que la calidad de la información sigue bajando si no existe atención alguna por parte del personal capacitado que la maneja.

Tradicionalmente, una vez comprobado, que los datos hayan sido capturados o transcritos correctamente dentro de los sistemas, ya nadie se fija en el grado de calidad que tienen o mantienen. Una de las razones más comunes para esto es la falta de una arquitectura, que permite que los datos sean documentados con cierto orden. La creación de una consciencia de trabajo en equipo y enfocada en la calidad es indispensable, para asegurar el éxito empresarial en la actualidad.

El éxito o el fracaso de una empresa depende en gran parte del uso efectivo de los datos que deben de representar el mundo real, para poder proporcionar la información necesaria. Pocos negocios tienen la información que necesitan, a pesar de que tienen todos los datos necesarios.

Cada vez es más común tratar la información como un producto. Esto es una tendencia nueva que refleja la necesidad para que el beneficio obtenido de la información sea máximo.
La información tiene el mismo camino como cualquier producto. En su ciclo de vida existen procesos que la crean y mantienen y al final hay usuarios finales y procesos que la usan. A los productos de manufactura, se le esta vigilando permanente la calidad durante el proceso de producción. La información se empieza a tratar de la misma manera. Su calidad debería ser vigilada permanentemente durante todo el recorrido que esta tiene en la empresa.

En el capítulo uno se hace una descripción básica de los tipos de datos desde el punto de vista del negocio. Se menciona la importancia de los datos como materia prima, para obtener información. Se define la calidad de los datos así como la problemática generada por tener mala calidad, redundancia e incoherencia entre ellos.

En el capítulo dos se abordarán los problemas de la calidad desde diferentes enfoques, con el objetivo de entender su importancia. Se muestra que la calidad de la información es directamente relacionada con el servicio al cliente. El capítulo también trata las posibles soluciones para este tipo de problemas de la calidad de información Se enfoca a la importancia de tener un equipo de trabajadores capacitados y el consentimiento de la gente involucrada en el proceso de trabajo. Se proporcionará una clasificación actual de herramientas y servicios para la limpieza y reingeniería de la información que proveen las distintas empresas del mercado.

En el capítulo tres se explica la problemática de una empresa contemporánea, la cual debido a la rápida penetración de la tecnología nueva y a la mezcla con tecnologías hereditarias, experimenta uno de los problemas de la calidad de la información – falta de integración en la información sobre sus clientes. Se explican la problemática y se describe el concepto de la solución con la que llegaremos a tener una mejor calidad de la información. Se proporciona un algoritmo para integración de los clientes de esta empresa.

En la actualidad, la importancia y la complejidad de los problemas de la calidad de la información llevan invariablemente a las empresas hacia la necesidad de tener y seguir una estrategia para su resolución.

CAPITULO 1
LA IMPORTANCIA DE LOS DATOS COMO MATERIA PRIMA PARA OBTENER INFORMACION CON CALIDAD

1.1 ANTECEDENTES EN EL MANEJO DE LOS DATOS Y LA INFORMACION

Desde los inicios de la era de la computación, la calidad de los datos y la información no siempre se consideraba como un aspecto importante en el desarrollo de los sistemas automatizados. En el principio los datos se manejaban por los programadores. Se introducían por los capturistas y pasaban por validaciones, pero el usuario final solo recibía los reportes. Con la introducción de los terminales, los usuarios empezaron a tener acceso para introducir y consultar la información.

En cuanto a la manipulación, el almacenamiento, la búsqueda y el acceso de la misma, todo se desarrollaba en los archivos secuenciales, donde los registros se clasificaban ya sea en orden ascendente o descendente basándose en la llave primaria. Los registros se graban en un archivo, uno detrás del otro, basándose en el orden secuencial lógico de la llave primaria. La limitación más importante en este tipo de organización secuencial consiste en que para poder tener acceso a un registro en particular, la búsqueda debe de leer todo el archivo desde el principio. En promedio, se tiene que leer la mitad del archivo antes de encontrar el registro. Esto consume mucho tiempo y más todavía en los archivos grandes. En esta época el problema de la calidad todavía no sobresaltaba con tanta fuerza. Los datos se introducían por los capturistas, pasaban por validaciones y se manipulaban por los programadores o personal especializado. El usuario final sólo tenía acceso a la información a través de los reportes emitidos. Después de la creación de los dispositivos de acceso directo, aparece el método ISAM de acceso con ayuda de índices (Index Secuential Acces Method). Los registros en un archivo ISAM se almacenan en una pista uno detrás del otro, dentro de un cilindro, en orden ascendente por la llave primaria hasta que la pista esta totalmente ocupada. Los registros subsecuentes, se mandan a la siguiente pista del mismo cilindro, por lo que no es necesario el movimiento del brazo de lectura/escritura y el tiempo de búsqueda es mucho menor. En el caso de ISAM los datos son separados entre diferentes archivos, unos están en archivo de datos y otros están en archivo de índices. Estos dos archivos tienen que ser sincronizados, si se pierde la sincronización entre ellos se pierde la información. En esta etapa en primer plano sobresale la integridad física de la información y los creación de nuevas formas de búsqueda rápida de información. La integridad lógica de la información y la calidad estaban en segundo plano.

Para mejorar las insuficiencias de estos métodos se crea un nuevo método de acceso llamado VSAM. Existen tres métodos de acceso a los datos. En todos ellos los registros se diseccionan internamente no por sus direcciones físicas, sino por medio de sus desplazamientos relativos desde el origen del archivo. De esta manera, los archivos VSAM alcanzan cierto grado de independencia del dispositivo donde se almacenan. Los bloques de un sistema de este tipo son llamados intervalos de control, que es un área contigua de almacenamiento con longitud fija. En un archivo existen diversos intervalos de control. El tamaño de estos intervalos varía de archivo a archivo y es independiente a las unidades de almacenamiento. Con esto se logra por primera vez que el archivo controle su propia integridad física de manera que no se puedan perder los datos. Esto es el primer indicio de integridad de información pero a nivel físico. Durante esta época, con la introducción de los terminales, los usuarios empezaron a tener acceso para introducir y consultar la

información. De aquí surge otro aspecto importante para la calidad - la integridad lógica de los datos y la información. Sus inicios empiezan con las bases de datos. Esto significa que los datos tienen que ser coherentes. Las primeras bases de datos son de tipo red y de tipo jerárquicos. En el tipo de red se puede ligar cualquier archivo por medio de un índice. Pero sin duda mayor impacto tuvieron las bases da datos jerárquicas que aparecieron a finales de la década de los años sesenta. El tipo jerárquico es cuando la estructura es de tipo árbol y se puede mover por las ramas, para acsesar la información. No se puede accesar información de manera horizontal. Por otro lado proporcionan facilidades de indexación para segmentos en cualquier nivel jerárquico y que se les pueda accesar directamente vía sus claves secundarias. Para hacerlos capases de manejar estructuras de bases de datos parecidas a redes, se usan las relaciones lógicas para vincular segmentos en la misma o en otras bases de datos. Las bases de datos jerárquicas no proporcionan a los usuarios la flexibilidad de crear escenarios distintos a consultas con propósitos predefinidos. Una de las ventajas que tienen sobre los sistemas relacionales es su excepcional rendimiento en bases de datos grandes, especialmente cuando el volumen de transacción es mayor. La razón de esto es que los segmentos pertenecientes a una ocurrencia de la raíz se pueden almacenar en un solo registro, mientras que en el sistema relacional cada registro conceptual se guarda como un archivo por separado. No obstante las bases de datos jerárquicas empezaron a ceder el paso a las bases relacionales.

Los sistemas de bases de datos relacionales estuvieron disponibles a principios de los años ochenta, una década después de los sistemas de red y jerárquico. El enfoque relacional es sustancialmente distinto a otros enfoques en términos de sus estructuras lógicas y del modo de las operaciones de entrada/salida. En el enfoque relacional los datos se organizan en tablas llamadas relaciones, cada una de las cuales se implanta como un archivo. En terminología relacional una fila es una relación y representa un registro o una entidad, cada columna en una relación representa un campo o un atributo. Así una relación se compone de una colección de entidades o registros cuyos propietarios están descritos por cierto numero de atributos predeterminados implantados como campos. La relación entre dos archivos se establece implícitamente por la presencia de un campo común en ambos archivos. Una de las características del enfoque relacional es la simplicidad de su representación lógica. Para el usuario final es mas fácil entender las tablas que los complejos arboles o estructuras de red. En el modelo relacional una tabla es una entidad separada que no ocupa un nivel jerárquico fijo. En el modelo de datos no se describen relaciones padre - hijo o propietario miembro. Se puede accesar cualquier archivo de manera directa sin la necesidad de navegar por la red de varios niveles desde la raíz. Esta facilidad de enfoque da como resultado el acceso a los datos sin depender de la trayectoria así como una mayor independencia de los mismos.

La consecuencia es que en este tipo de bases de datos ya están introducidos los primeros principios de calidad de información y la integridad relacional. Se introduce el diccionario de metadatos que asegura la precisión, integridad y completes de los datos La integridad referencial puede ser por definición o programada. Por definición es cuando se definen relaciones entre llaves primarias y secundarias. El caso programable se hace por medio de disparadores que también son llamados trigerrs. Él controla la modificación de tablas por medio de solo tres comandos, insertar, borrar y actualizar. Esto ayuda a tener una integridad mas sofisticada y no únicamente con base de las llaves primarias y secundarias. En cuanto a

los datos se introduce el concepto de datos por omisión donde se especifica que valor puede tomar un dato en caso de que falte su valor. El concepto de regla también se introduce y lo que hace es definir el dominio de los valores de los datos. También existen funciones ya integradas para la manipulación de los datos. Así también se introduce la seguridad de acceso, se crean nuevas herramientas amigables, para una mejor representación y manipulación de la información por parte del usuario final. Ahí se introduce el concepto de integridad distribuida donde se tienen que asegurar la integridad referenciál, no nada mas dentro de una base de datos, sino entre dos o más de ellas situadas en diferentes sitios físicos. Se implementan transacciones distribuidas que terminan cuando el proceso se acaba correctamente en todos los nodos donde existen bases de datos. Si en alguna base la transacción falla, se tiene que efectuar reverso de la transacción en todos los nodos.

Los años ochentas han sido anunciados como una época de auge en la Información. Se revelo el poderío computacional en los sistemas grandes y aun más a nivel estación de trabajo individual. Para muchas grandes organizaciones la PC no llevo a la resolución de todos los problemas. Los datos frecuentemente se escribían y reescribian manualmente, los errores entraron poco a poco y las decisiones administrativas a veces se basaban en información inexacta. El objetivo por lo tanto, fue encontrar la manera de proporcionar un método sencillo para permitir que se canalicen la necesidades de calidad de negocio de la gran comunidad de usuarios finales.

Después de las bases de datos relacionales el siguiente nivel en cuanto al manejo de la información y su calidad es el data warehause. Su existencia es necesaria porque con las bases de datos relacionales en las empresas empiezan a existir muchas bases diferentes y la coordinación entre ellas es cada vez, más difícil. Por lo tanto las empresas necesitan una globalizacion de la información que tienen almacenada. Un data warehouse es una colección de software y datos organizados para la limpieza, transformación, y almacenamiento de los datos desde una gran variedad de recursos informáticos. Analiza y presenta la información, participa de manera activa en el proceso de toma de decisiones y los procesos tácticos y estratégicos de la empresa. Esto se logra por medio de auditorias y controles, que son una colección de chequeos y balances que aseguran, la extracción, limpieza, sumarización, transformación y cargado de los procesos y su desempeño correcto. Los datos se extraen desde el recurso indicado, son limpiados, sumarizados y transformados y se cargan en un archivo especificado. Esto se hace por medio de un almacenamiento completo, único y consistente y presenta la información a los usuarios finales, para que puedan entenderla y usarla en el marco de la empresa. El data warehouse es un paso hacia un mejor orden en los procesos relacionados con la calidad de los datos y la información.

1.2. TIPOS DE DATOS PARA EL MANEJO DEL NEGOCIO

Para una mayor comodidad en el contexto de este trabajo, se define dato como la representación computarizada de la información del negocio. Los datos basados en computadoras se han usado por mucho tiempo para ejecutar y administrar el negocio. Tales datos, llamados datos empresariales, representan la situación del negocio y su valor radica en el significado que representa. El dato se puede considerar como producto, porque se produce, se compra y se vende de la misma manera en que se hace con cualquier producto físico. Ejemplo de ello son las películas digitalizadas y los libros. Los datos pueden

clasificarse de diferentes maneras, según la necesidad, pero resulta muy complicado elaborar una clasificación general de los datos, porque esta depende del significado, de la interpretación y del punto de vista. Unos ejemplos de la complejidad y la variedad de los criterios que se pueden aplicar, para obtener una clasificación se muestran a continuación. Por ejemplo, los datos pueden ser personales, en donde su dueño los puede cambiar como le parezca o públicos, en donde su uso se comparte entre un gran número de personas y donde cualquier cambio requiere de un cuidadoso manejo. Por otro lado si los miramos desde el punto de vista de la procedencia, los datos pueden ser internos o externos, según el lugar donde se originan. Como su nombre lo indica los datos internos son todos los datos del negocio que están dentro de la empresa. Los datos externos son útiles y requeridos, pero no se encuentran en la institución, por lo que por medio de diferentes caminos se consiguen e introducen, para apoyar el proceso de trabajo. En el pasado la mayoría de los datos de interés, para una organización se originaban dentro de la misma. Aun cuando los datos eran originados externamente, el número de fuentes era lo suficientemente pequeño y los volúmenes de datos lo suficientemente bajos para que el impacto de los datos externos en la arquitectura global, sea muy significante. El crecimiento extraordinario de Internet en los últimos años ha causado un crecimiento exponencial en los volúmenes de los datos capturados electrónicamente y de los datos de acceso libre, que dejan todas las organizaciones En este trabajo nos enfocamos principalmente los datos internos. Ahora, si por otro lado consideramos el ciclo de vida de los datos, esto nos llevará a la circulación del dato. La circulación de los datos refleja el lugar donde se encuentran posicionados sobre la línea del flujo de los procesos de la empresa. Esta consideración es otro punto de vista para una clasificación.

1.2.1. CRITERIOS DE CLASIFICACION

Los datos empresariales son los datos requeridos para operar y administrar el negocio. Ellos representan la actividad que el negocio emprende y los objetos en el mundo real, clientes lugares y productos con los que se trata. Son creados y usados a través de sistemas de procesamiento de transacciones y sistemas de soporte de decisiones. Dentro de los datos empresariales, se deben manejar diferentes tipos de datos. Los criterios mostrados a continuación se basan en una valoración de cómo los negocios utilizan los datos. Se escoge este criterio no solamente como consideración teórica, sino también por la experiencia de lo que se ha trabajado en las implementaciones de administración de datos. Los tipos de datos que surgen basados en este criterio se usan después para determinar la colocación de los datos, su nivel de duplicación y las reglas para manejarlos.
En este trabajo vamos a definir cuatro criterios empleados para determinar los tipos de datos empresariales.

El primer criterio es que los datos siempre tienen un significado o que son representaciones de algo que tiene significado. Los trabajadores capacitados usan el significado para interpretar la información de forma correcta. Si él significado no esta definido, entonces la interpretación de los datos es subjetiva.
Como segundo criterio se considera que los datos pueden estructurarse de una manera compleja, consistiendo en muchos campos bien definidos y relacionados entre si por el significado común. Los sistemas de información de la administración se han enfocado a los datos bien estructurados, tales datos comúnmente tienen las siguientes características; una

proporción importante de los datos es numérica, existen múltiples atributos para cada entidad expresados como múltiples campos por registros o múltiples columnas por tabla y casi siempre hay múltiples relaciones entre las diferentes entidades. Al otro extremo de la clasificación se encuentran los datos no estructurados, características de los cuales son las opuestas de las listadas anteriormente. Imágenes, audio, video son ejemplos de datos altamente no estructurados. Los datos textuales como notas y documentos caen entre los dos extremos. La importancia entre los tipos de datos menos estructurados esta creciendo rápidamente en todas las empresas y consecuentemente en los sistemas de información. Los datos no estructurados tienen menor oportunidad de ser clasificados de acuerdo a los tipos de datos definidos anteriormente. No obstante, es mucho más apropiado clasificar los datos no estructurados de acuerdo a las categorías de datos que ya se han definido que al usar un esquema totalmente diferente. La racionalización para esta metodología se funda en el reconocimiento de que ambos datos, los estructurados y los no estructurados, son usados juntos en los mismos procesos por los usuarios y deben ser manejados en la misma extensión por los departamentos de sistemas de información. Por consiguiente arquitectónicamente es mejor concentrarse en las similitudes más que en las diferencias entre los datos estructurados y los no estructurados.

El tercer criterio es el desempeño o uso en la empresa, los datos se usan para cubrir dos objetivos principales: Los datos operacionales se utilizan para ejecutar el negocio y se relacionan con las acciones o decisiones a corto plazo. Los datos operacionales son los datos primarios de la empresa dentro de la organización y son la fuente de todos los datos informáticos. Los datos informáticos se usan para administrar la empresa en un plazo más largo. Ambos, los datos operacionales e informáticos se estructuran de acuerdo a su acceso y a su necesidad de uso. El cuarto criterio es su alcance. Los datos pueden representar un elemento sencillo o una transacción, o pueden ser una suma del efecto neto de un juego de elementos o transacciones: los datos detallados o datos atómicos son críticos para ejecutar el negocio, pero también se usan en alguna de las tareas más sencillas de administración dentro de la empresa. A menudo se enfocan a objetos o transacciones básicas tales como productos individuales, ordenes o clientes. Los datos sumarizados se usan en el manejo de la empresa y para mostrar una vista amplia de la manera en que la empresa esta operando. El significado de estos últimos dos tipos de datos se menciona también en la parte de datos derivados.

1.2.2. TIPOS DE DATOS EMPRESARIALES

La figura 1.1 representa los datos empresariales agrupados y clasificados según el uso que se le da en la empresa. Los procesos operativos son el conjunto de la actividad manual mas la actividad de los sistemas operacionales de cómputo. Ellos involucran los datos de tiempo real y los datos conciliados y también pueden usar datos externos que provienen de otros organismos. Los procesos que soportan la toma de decisiones requieren de datos combinados de muchos procesos y muchas identidades o tablas normalizadas. Este proceso de unión requiere un enfoque bastante formal para asegurar su integridad, validez y calidad. *Este enfoque se realiza por los sistemas informativos. Los datos de capas (conciliados y* derivados) cubren a la organización completa y son de un alcance más extenso, que al que los usuarios finales puedan recurrir. Estas razones de negocio limitan el uso directo total de la capa de datos conciliados para los propósitos de manejo de la información por parte del

usuario final. También, las funciones requeridas, la unión de tablas y la selección de pequeños subconjunto de datos son de un valor monetario elevado. El acceso a esta capa de datos conciliados es limitado a un pequeño numero de analistas y trabajadores capacitados, que tienen que ver el negocio como un todo. La vasta mayoría de los usuarios finales persigue sus necesidades de negocio a través de la capa de datos derivados.

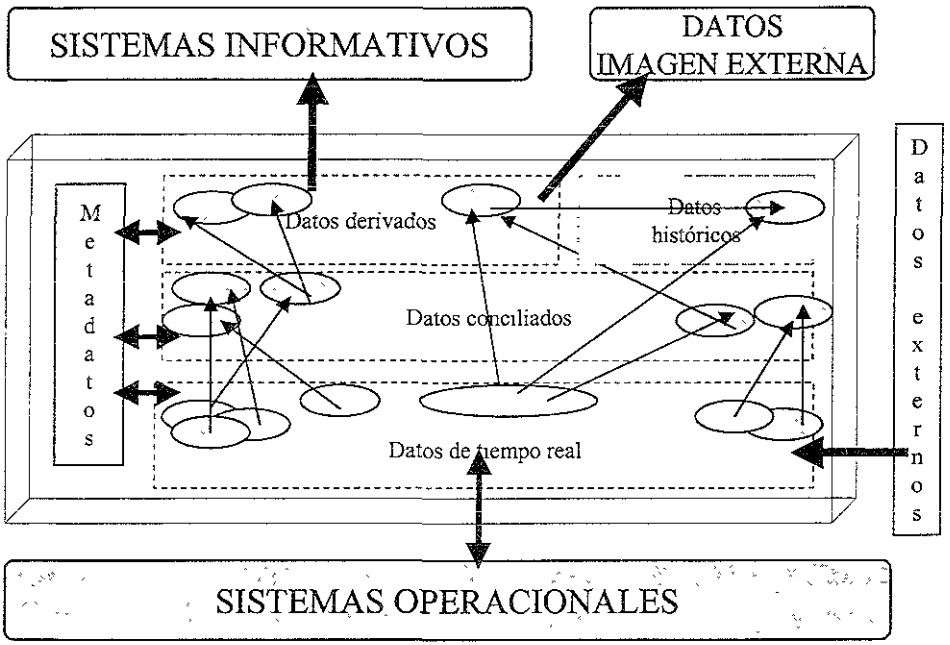

Figura 1.1 Los datos empresariales y las relaciones entre ellos

De esta manera, la capa de datos derivados consiste en conjuntos de datos que han sido optimizados para las necesidades de algunos departamentos en particular, algún grupo de usuarios, o incluso individuos. Una razón para que sea útil de separar los datos conciliados con los datos derivados es que dichas necesidades de información de manejo son ampliamente predefinidas y repetitivas. La implementación técnica de esta simple observación produce una dramática reducción en las necesidades de recurso de computo.

1.2.2.1. DATOS DE TIEMPO REAL

Los datos de tiempo real son los datos del presente o de hasta el instante que representan el estado presente de la empresa y se utilizan para ejecutar el negocio. Ocurren a un nivel detallado y se acceden en un modo lectura/ escritura.

En general, los datos de tiempo real son los datos creados, manipulados y utilizados por las aplicaciones operacionales o de producción. Tradicionalmente encontramos estos datos en archivos o bases de datos en ambientes de mainframe y son controlados y administrados

por el departamento de sistemas de información. Aunque las nuevas bases de datos puedan ser relacionales, incluso hoy en día la mayoría de estos datos aun existen fuera del mundo de las bases de datos relacionales. En sistemas más antiguos, tales datos se encuentren pobremente estructurados o contienen estructuras muy complejas provocadas por su repetido mantenimiento durante el paso del tiempo.

Estos datos de tiempo real no se encuentran únicamente en aplicaciones de mainframe o aplicaciones de legado. Una nueva categoría de aplicaciones cliente - servidor crea datos de tiempo real en estaciones de trabajo y servidores. Los datos de tiempo real se distribuyen a lo largo de la empresa y están raramente bajo el control directo de los departamentos de sistemas de información. Cada vez más los datos de tiempo real se originan fuera de la compañía entera. Esto ocurre cuando los procesos inter - empresariales tales como ordenes y facturas se transfieren entre las organizaciones usando el intercambio electrónico de datos, y los datos que ingresan se emplean como la base de las actividades receptoras de la compañía.

Algunos ejemplos comunes de datos de tiempo real son la base de datos de clientes, los registros de llamadas, y las cuentas contables. Tales datos existen en distintos formatos y localizaciones. Su tamaño varía enormemente. Su factor común es el uso que se les da para ejecutar el negocio a un nivel detallado. Sin embargo los datos de tiempo real son la fuente fundamental de todos los demás datos empresariales.

Los datos no estructurados de tiempo real corresponden a las imágenes electrónicas de transacciones de la empresa que no se pueden descomponer fácilmente en muchos campos de datos discretos. Los sistemas de procesamiento de producción de imágenes son buenos ejemplos de aplicaciones operacionales que crean y administran tales datos. De esta manera, la imagen de una forma de derecho de seguro, una copia almacenada de estado de cuenta impresa, o la imagen de una licencia de conducir con firma y fotografía son datos no estructurados de tiempo real. Las notas en un sistema de correo electrónico son un ejemplo más avanzado de esa clase de datos.

1.2.2.2. DATOS DERIVADOS

Los datos derivados son aquellos que se derivan a través de ciertos procesos, a partir de los datos de tiempo real. Se usan para manejar la empresa en el modo de únicamente lectura, más que en alguna operación diaria de la empresa. Pueden estar a un nivel detallado o resumido. Ya que se derivan de los datos de tiempo real, son ya sea de una naturaleza de "punto en el tiempo", representando una vista del negocio en un momento determinado, o de naturaleza periódica, preservando un registro histórico del negocio sobre un periodo de tiempo.

Los datos derivados conforman aquellos conjuntos de datos que se han utilizado tradicionalmente para el soporte de tomas de decisiones. Hoy en día se encuentran por toda la organización, desde las bases de datos relacionales, hasta en paquetes de hoja de cálculo especializados, en las PCs, en intranet y todo lo que se encuentra entre ellos. Aunque es un ideal que el proceso de derivación de los datos debe de estar automatizado, en algunos

casos dicho proceso todavía puede ser manual con los contenidos de reportes impresos que fueron tecleados dentro de las herramientas de información.

El punto hasta el cual los datos derivados difieren de su fuente depende de los requerimientos de la empresa. En los casos más sencillos los datos derivados pueden ser de toma instantánea o de copia de punto en el tiempo de los datos de tiempo real y de esta manera se encuentran a un nivel detallado.

Otro tipo importante de derivación es la sumarización, en donde los datos detallados se traen a un nivel más alto de agregación. Ambos, los datos detallados y los sumarizados pueden ser subseriados seleccionado únicamente algunos de los campos (subseriación vertical) ó algunas de las instancias de los datos (subseriación horizontal). Los resúmenes de venta, por ejemplo, pueden contener únicamente cifras de ganancias netas y pueden calcularse solamente para regiones específicas.

Finalmente pueden derivarse nuevos datos de una combinación de datos o registros existentes. Este tipo de derivación a menudo llamada "de enriquecimiento", es uno de los tipos más eficientes de derivación, para proporcionar vistas nuevas y originales de la empresa. Sin embargo es también diferente en cierto modo con respecto de los otros datos y de procesos de derivación mientras que los otros procesos son inherentemente auto-consistentes. El enriquecimiento es consistente únicamente si los datos que se están combinando se relacionan lógicamente unos con otros. Muchos de los problemas que los usuarios y los departamentos de sistemas de información enfrentan con los datos derivados, son el resultado de combinar datos que no están estructurados lógicamente.

Los datos no estructurados derivados pueden considerase como sumarizaciones o abstracciones de los datos de tiempo real, igual que para los datos estructurados. Sin embargo el proceso de sumarización es mucho menos mecánico para los datos no estructurados que para los datos estructurados. Las cifras de ventas (datos estructurados) pueden sumarizarse por región de ventas simplemente añadiendo las entradas de detalle a las entidades; la sumarización de los comentarios de satisfacción de los clientes por tipo de producto requiere de un gerente de producto que lea los datos detallados y elabore un reporte de resumen.

1.2.2.3. DATOS CONCILIADOS

Los datos conciliados son generados por un proceso diseñado para asegurar una consistencia interna de los datos resultantes. Este proceso opera en datos de tiempo-real a un nivel detallado. Un segundo aspecto clave de este proceso de generación es que mantiene o crea una serie de datos históricos. Es por esto, que los datos conciliados deben verse como una categoría especial de datos derivados.

El paso de la conciliación de datos, puede tomar datos de diversos sistemas heterogéneos geográficamente distribuidos y los combina y mejora en una imagen simple y lógica del modelo de datos empresariales. El propósito de los datos conciliados es ser fuente autoritaria y única de todos los datos que requieren los usuarios finales de administración de información o sistemas de tomas de decisiones. A partir de estos datos se puede derivar

cualquier combinación de datos que se requiere por los usuarios finales y en el momento que lo requieran.

Se concilian los conjuntos de datos en la capa de datos de tiempo real entre si como parte del proceso de ser copiados a la capa de datos conciliados. Este paso se lleva a cabo por la necesidad de depuración de los datos de tiempo real, para eliminar sus inconsistencias e irregularidades. En este paso no se crean nuevos datos, el valor agregado viene de la reconciliación misma.

En ambientes tradicionales de soporte de decisiones, rara vez se definen los datos conciliados explícitamente. En muchos casos, no existen en absoluto. Cuando existen rara vez se almacenan físicamente, siendo únicamente el resultado lógico de ciertas operaciones que tienen lugar en el proceso de derivación. En otros casos, solamente existen archivos temporales. De esta manera no parecen tener ninguna consecuencia dentro de la empresa. De hecho, los datos conciliados son el elemento pivote de un data warehouse. Como resultado de la una metodología de desarrollo, dirigida a aplicaciones, los datos de tiempo real no son autoconsistentes sobre el alcance total de la empresa. Esto hace a la conciliación de datos una necesidad.

Así cuando los datos de fuentes múltiples tienen que combinarse, los desarrolladores primero deben analizar la estructura y contenido de las fuentes para definir las reglas de combinación. Después necesitan desarrollar un proceso que dé fuerza a estas reglas. Típicamente dichos procesos incluyen funciones tales como igualación y manipulación de los campos, conversión de los contenidos de los campos a formas consistentes y en situaciones extremas, varios tipos de corrección de errores.

Los siguientes ejemplos pueden ilustrar mejor el concepto de conciliación. Las llaves de identificación única para identidades tales como clientes o productos a menudo difieren de aplicación a aplicación. En algunos casos, puede existir una relación directa entre ambas estructuras. Por ejemplo, una aplicación antigua utiliza un código de seis dígitos como llave para cliente, mientras que una aplicación más reciente ha extendido este campo a ocho dígitos, pero utiliza el formato anterior de seis dígitos como una subserie de la nueva definición. Combinar estos dos campos es un asunto tan simple como el de extender el campo recibido de la aplicación más antigua. Un problema más complejo surge cuando no hay relación directa entre las dos llaves. Aquí puede hacerse necesario construir manualmente una tabla de referencia usada para conciliar estas dos series de datos o aplicar procesos de integración de los clientes duplicados.

Los campos codificados a menudo requieren conciliación. Dos ejemplos clásicos son los códigos "M/F" o "1/2" para representar el sexo masculino o femenino en diferentes aplicaciones y la variedad de sistemas de codificación por país que se usan en los distintos departamentos. La conciliación en estos casos requiere de la definición de una serie de valores comunes y de traducciones de los valores de base diferentes.

Un tipo de conciliación más complejo se relaciona a la dependencia de tiempo de los datos en diferentes aplicaciones. Las aplicaciones financieras producen datos que se basan en la posición en tiempo al final del mes. Si estos datos han de combinarse con los datos de

venta con la llave al cierre del negocio diario, entonces se requiere de un paso de conciliación que convierta los datos financieros a datos de venta de consistencia de tiempo. Este proceso de conciliación diferirá de acuerdo al día del mes, variando desde "sin acción en absoluto" al final del mes, hasta una aplicación de " cambios mensuales totales" el día anterior al cierre del mes.

Cualquier conciliación requerida de datos no estructurados ocurre a través de su asociación con los datos estructurados. Así una cuenta textual de un accidente de tráfico guardado en un sistema de procesamiento de textos puede conciliarse con un vídeo de la escena del accidente guardado en una base da datos. Ambos son ejemplos claros de dato son estructurados. Por lo tanto uno puede concluir que los datos no estructurados conciliados no tienen una existencia física.

1.2.2.4. DATOS HISTORICOS

Los datos históricos juegan un papel importante en la tendencia del análisis, o para patrones de compra y uso, los cuales se enfocan en áreas particulares de los datos empresariales. Los datos históricos también constituyen un componente grande e importante de los datos de calidad empresarial, ya que proporcionan un registro definitivo del negocio. Los requerimientos para mantener un registro histórico caen dentro de dos grandes áreas:
Una es la vista del negocio en un momento dado y por otro lado el análisis de las tendencias del negocio.
En general, los usuarios finales necesitan dar vistas al negocio tal como existe en diferentes tiempos. Existen ciertos momentos en la empresa que tienen una significancia particular en el negocio. El cierre de contabilidad o los periódicos de pago de impuestos, y eventos mayores de negocio tales como reorganizaciones o adquisiciones son tiempos que requieren un análisis mas profundo. Estos análisis y vistas tienen que ser estables permitiendo la misma búsqueda en diferentes tiempos para producir los mismos resultados. La necesidad para algunas de estas vistas del negocio se conoce por adelantado, para que el departamento de sistemas de información pueda, por lo tanto, pre-almacenar los datos requeridos. Otras necesidades no son predecibles, ni en el tiempo, ni en el contenido de datos. En tales casos los datos no pueden almacenare por adelantado, así que los usuarios finales necesitan un método que posibilite la generación de estos datos retrospectivos.

La otra área es el análisis de tendencias de negocio que es apropiado para establecer el nivel fundamental del cambio del negocio – la transacción del negocio como su base para este análisis. Los datos que contienen un registro de transacciones de negocio sobre un periodo de tiempo ya han sido identificados como datos históricos. Los datos históricos pueden pertenecer a otro tipo de datos si sus características se lo permiten. La extensión del tiempo de estos datos históricos depende del negocio en que la compañía esta involucrada. Hay que recordar que en consecuencia de no almacenar todos los datos históricos al nivel de detalle necesario, será imposible cumplir con algunos requerimientos de análisis históricos futuros, porque ya no habría algunos datos disponibles.

1.2.2.5. DATOS REPLICADOS

El copiado de datos podría describirse como la profesión más antigua en el mundo del procesamiento de datos. Se han creado copias de datos desde los primeros tiempos. Dependiendo del propósito de los datos copiados, la copia podría haber sido idéntica a su fuente o podría ser cambiada de alguna manera específica. Sin embargo la replicación va más allá del copiado. Tradicionalmente el copiado ha sido una actividad conducida por necesidades inmediatas de datos, iniciada y diseñada sin previsión a consecuencias más amplias. La replicación de datos es copiar bajo control. La necesidad de las réplicas es evidente y satisface las necesidades de los datos de tiempo real, conciliados, y derivados.

1.2.2.6. METADATOS

A medida que la variedad de datos almacenados y utilizados en una empresa aumenta, y a medida que se expande la diversidad del uso de los datos, surge una necesidad para formalizar una manera de describir los datos y para asegurar su uso consistente y completo. Los metadatos son datos acerca de otros datos. Sin embargo esta definición implica que la materia de los metadatos son únicamente los datos. En realidad los datos empresariales no existen en un vacío. Son creados, mantenidos, y accesados a través de procesos de negocio que se implementan por medio de aplicaciones. Por lo tanto la empresa necesita una descripción completa de sus datos empresariales y los procesos que les dan mantenimiento y los usan. De este modo los metadatos describen una gran cantidad de aspectos de la empresa y de las funciones de aplicación correspondientes. Los Metadatos raramente entran o salen de la organización como un fin en si mismos. Mas bien acompañan a los datos del negocio a través de las fronteras de la organización. Los metadatos son parte indispensable de la estructura de un data warehouse.

El data warehouse es un paso adelante en el proceso de crear orden y mejorar la calidad de los datos. La definición se basa en la historia de la informática enfocada al usuario final. Un Data Warehouse es simplemente un almacenamiento único, completo y consistente de datos obtenidos de una variedad de fuentes, los cuales pone a disponibilidad de los usuarios finales de tal manera que puedan entenderse y usarse en un contexto empresarial.
Alcanzar la integridad y consistencia de los datos en el ambiente de los sistemas de información hoy en día resulta, no obstante, lejos de ser simple y requiere de un esfuerzo en el ámbito de toda la empresa.

Los criterios mencionados arriba permite determinar el lugar de cada dato, su significado y alcance en el complejo proceso de negocio. A continuación veremos algunas consideraciones para la calidad de los datos.

1.3. CALIDAD DE LOS DIFETENTES TIPOS DE DATOS

1.3.1. INTEGRIDAD DE LOS DATOS

El primer componente de la calidad de datos es la integridad. Tener datos íntegros consiste en aplicar varias técnicas que sirven para determinar la manera en la que los datos están mantenidos dentro de los recursos informáticos. Estas técnicas se aplican para que el

usuario se asegure que los recursos informáticos tienen una alta integridad. Se pueden distinguir las técnicas principales que incluyen: integridad de los valores, integridad estructural e integridad de retención.

1.3.1.1. INTEGRIDAD DE LOS VALORES DE LOS DATOS

La integridad de los valores de los datos es un subconjunto de la integridad de los datos que especifican los valores permitidos para cada tipo de dato y para las relaciones que existen dentro de la arquitectura del sistema donde están. El valor integral de un dato es especificado como el valor actual que toma el dato, ó un valor codificado valido. La regla de la integridad de los datos, es una declaración que define y controla los valores actuales de los datos o sus valores codificados permitidos.

El valor integral de un dato puede ser especificado como característica única del dato ó como una relación entre características de varios datos.

Ejemplo 1.1: Existe una relación entre el tipo de empleado y su antigüedad. Ciertos tipos de empleados pueden únicamente convivir con cierto tipo de antigüedad. La integridad de los valores para la relación entre tipo de empleado y su antigüedad está definida en las primeras dos columnas de la tabla. Se pueden añadir columnas adicionales como la fecha de ingreso y la fecha de salida, o con cualquier otra característica relacionada con la dependencia entre el tipo de empleado y su antigüedad. Los datos guardados en fecha de ingreso y fecha de salida son útiles cuando sus valores o las relaciones entre ellos cambian en algún momento. (Ver tabla 1.1)

Ejemplo 1.2:
La regla de la integridad puede ser especificada por el tipo de dato. Por ejemplo se puede determinar que para representar la edad de un empleado se van a usar tres dígitos y para el caso del nombre una regla específica sería de treinta y cuatro caracteres alfabéticos, alineados a la izquierda.

Tabla 1.1. Relación entre el tipo de empleado y su antigüedad.

Tipo del empleado	Antigüedad del empleado	Fecha de ingreso	Fecha de salida
1	A	10/01/91	
1	B	10/01/91	
2	B	10/01/91	
2	C	10/01/91	
3	A	10/01/91	
3	B	10/01/91	
3	C	10/01/91	12/31/92

Ejemplo 1.3:
Las reglas de la integridad de los datos pueden ser escritas de una manera más formal. Por ejemplo si una empresa desea controlar la edad de sus trabajadores se puede usar la

siguiente regla. La empresa no quiere que ingrese gente mayor de cuarenta años, y su fecha de nacimiento tiene que ser menor que la fecha de hoy.
1960<= El año de la fecha de nacimiento del empleado <= Año actual

Ejemplo 1.4:
La regla de integridad de los datos para un registro federal de contribuyentes (RFC) depende del tipo de persona. (Ver tabla 1.2)

Tabla 1.2. Relación entre tipo de persona y el formato de RFC.

Tipo de persona	**Regla para el formato de RFC**
Persona moral (empresas)	[A-Z][A-Z][A-Z][0-9] [0-9] [0-9] [0-9] [0-9] [0-9]%
Persona física	[A-Z][A-Z][A-Z][A-Z][0-9] [0-9] [0-9] [0-9] [0-9] [0-9]%

Para el caso de personas morales el RFC determina como alfabéticas las primeras tres posiciones y como numéricas las siguientes seis. El signo de porcentaje al final significa que la regla se aplica para las primeras nueve posiciones para el caso de una persona moral y diez para el caso de una persona física. Las seis posiciones numéricas deben de corresponder a una fecha válida en el formato: dos posiciones para el año, dos posiciones para el mes del año y dos para el día del mes. Por ejemplo no puede existir un valor para el día mayor que 29 si el mes es 02 (febrero).

1.3.1.1.1 INTEGRIDAD CONDICIONAL DE LOS VALORES DE LOS DATOS

La integridad condicional de los valores de los datos especifica si los valores de los datos son mandatorios, opcionales ó se cumplen bajo ciertas condiciones (preventivos), se pueden usar notaciones como (M), (O), (P), para cada uno de los tres casos, ó también dos para mandatórios, cero o uno para opcionales y cero para preventivos, pero por lo general estas notaciones tienen menos valor para el usuario final y pueden ser omitidos. La tabla 1.3 es un ejemplo de este tipo.

Tabla 1.3. Tipos de valores, para la integridad condicional.

Estudiante	
Nombre del estudiante	Mandatorio
Fecha de nacimiento del estudiante	Mandatorio
Cliente	
Nombre del cliente	Mandatorio
Ingresos del cliente	Opcional
Vehículo	
Potencia del vehículo	
Vehículo motorizado	Mandatorio
Vehículo no motorizado	Preventivo

Ejemplo 1.5:
El nombre del estudiante y su fecha de nacimiento son mandatórios al igual que el nombre del cliente, pero los ingresos del cliente son opcionales. La potencia del vehículo es mandatória para vehículos motorizados, y es preventiva para vehículos no motorizados. Estas condiciones están especificadas en la tabla anterior.
La integridad condicional de los valores de los datos puede ser especificada como la regla de integridad condicional. Por ejemplo la potencia del vehículo, en el ejemplo anterior puede ser definida como una regla de integridad condicional de los valores de datos, por medio de una matriz.

Si el vehículo es motorizado	Entonces el valor de la potencia es mandatório
Si el vehículo no es motorizado	Entonces el valor de la potencia es preventivo

Ejemplo 1.6:
La integridad condicional se puede especificar por medio de una anotación más formal. Las características opcionales de los datos son mostradas conjuntamente, separados por una línea vertical.
Empleado = {Edad del empleado | Fecha de nacimiento del empleado}

Esta notación sirve para mostrar conjuntos de datos que son excluyentes uno para el otro, o se requiere la fecha de nacimiento o la edad.

Ejemplo 1.7:
Se puede dar el caso de que por un lado se requiera el fabricante o la marca de un vehículo o en caso contrario el número de motor, esto se puede expresar de la siguiente manera:

Vehículo = { Nombre del fabricante, Nombre del modelo | Numero del motor }

Cualquiera de estos ejemplos es aceptable para definir y documentar la integridad de los valores de un dato. El punto importante es recordar que la integridad de los valores de los datos tiene que ser definida claramente, de una manera comprensible y aceptable para el usuario y para el negocio.

1.3.1.1.2. VALORES DE LOS DATOS POR OMISION

Los valores por omisión se usan con frecuencia durante la captura y la edición de los datos. En el contexto de la calidad de los datos esta práctica es conveniente en la mayoría de las situaciones, pero no en todas. Por ejemplo para el caso de una aplicación de crédito de un banco, la fecha de hoy puede ser introducida para la fecha del inicio del crédito. El monto del límite del crédito puede ser introducido basándonos en la política del banco usando datos específicos de la misma aplicación Desafortunadamente en otras situaciones, los valores por omisión no pueden ser usados. Tal es el caso por ejemplo, cuando el promedio de un estudiante universitario, al ingresar a la universidad, no se conoce. Un valor por omisión no es apropiado ya que él puede ingresar como alumno regular o como becario, y en algunas escuelas existen más opciones que ofrecen facilidades de acuerdo con el promedio obtenido.

Los valores por default necesitan auditorias que sirven para asegurar que se mantenga la calidad de los datos. Desde este punto de vista la mejor práctica es no tener valores por default. En caso de que una organización considere usar estos valores, entonces se necesitaran auditorias permanentes para asegurarse que los valores por default siguen validos. En caso de que la auditoria muestre que los valores no son válidos, el proceso debe de ser cambiado para tener valores correctos. Este método de auto corrección para valores por default mejora la calidad de los datos.

1.3.1.2 DOMINIO DE LOS DATOS

La integridad de los datos es definida por sus características y forma parte dentro del conjunto integrado de los recursos informáticos. El dominio de un dato es una característica indispensable que contiene descripción de la integridad de los valores y descripción de las reglas de integridad de los datos. El dominio de los valores de los datos contiene un conjunto de los valores que un dato característico puede contener bajo condiciones especiales. El dominio puede ser una lista de valores que pueden ser continuos o inconexos o una combinación de valores de varias características relacionadas. También puede contener fecha de inicio y fin, determinando la fracción de tiempo por el cual son validos estos valores. El dominio de reglas de un dato es en realidad un dominio que contiene las reglas de integridad de este dato. Éstas a su vez también incluyen fechas de inicio y fin que determinan la fracción de tiempo por el que son validas estas reglas. Cada dato característico debe tener su dominio correspondiente que especifica sus valores permisibles.

Ejemplo 1.8: Dentro de los ejemplos anteriores vimos como se definen los dominios de los datos del estudiante, del empleado, del cliente y del vehículo. Fue definida la integridad de los datos del estudiante, la integridad de la relación empleado - antigüedad, la integridad de los datos del cliente y la integridad de los datos del vehículo.

Si un dato tiene múltiples características que varían dentro de su característica principal, cada una de estas variantes especificadas debe tener su propio dominio. Una de las maneras más fáciles para determinar si existen variaciones en las características de un dato es revisar primero su dominio. En caso de que exista más de un dominio, entonces existe más de una variación característica del dato.

En el modelo lógico de un dato, un atributo representa una variante de la característica que tiene su dominio correspondiente. La principal razón de la necesidad de identificar las variantes de las características de los datos y sus dominios correspondientes es para asegurarse que cuando una variación de las características del dato sea escogida, dentro de la arquitectura, para ser atributo en un modelo lógico, debe de tener su dominio formal.

Por lo tanto cada característica del dato debe de tener su dominio.

La integridad del dominio de los datos es una parte importante para tener constantemente calidad en los recursos de datos integrados.

La integridad de los valores de los datos debe de ser progresivamente definida hasta que se obtenga la definición de la regla de integridad final. La regla de la integridad resultante es una cadena de texto de longitud no limitada. Esta técnica progresiva nos proporciona una combinación de reglas de integridad primarias, alrededor del recurso integral de datos. Estas reglas de datos primarias son desarrolladas usando las necesidades del negocio de una empresa y permiten responder a las mismas.

1.3.1.3 INTEGRIDAD DE LA ESTRUCTURA DE LOS DATOS

La integridad de la estructura de datos es la parte, que especifica la integridad de las relaciones entre los datos. Este tipo de integridad puede ser documentado por medio de un diagrama, una matriz ó una tabla.

Ejemplo 1.9: La existencia de un empleado implica la existencia de una persona, pero la existencia de una persona no implica la existencia de un empleado. De la misma manera la existencia de un cheque de pago implica la existencia de un empleado, pero el caso contrario no se cumple. (Ver tabla 1.4)
Una matriz de integridad de la estructura de datos, es cuando los datos se muestran por medio de una matriz. Aquí podemos usar el ejemplo anterior con los mismos valores. La matriz de integridad de la estructura de datos, no es gráfica como lo es el diagrama, pero puede mostrar la posibilidad de otras condiciones, como lo son por ejemplo las relaciones que existen entre estudiante y persona ó entre buen estudiante y persona. (Ver tabla 1.5.)
Los mismos datos son usados para crear una tabla de integridad de la estructura de datos. Esta tabla no es gráfica como el diagrama y no muestra otras posibilidades, como lo hace en el caso anterior la matriz, pero por otro lado tiene una ventaja que en algunos casos puede ser significativa porque requiere de menos espacio. En los dos primeros casos del diagrama y la matriz son generalmente usados para identificar la estructura de los datos integrados, mientras que la tabla de integridad se podría usar de manera más apropiada para documentación.

Tabla 1.4. Relación entre empleado, persona y cheque de pago.

	Empleado	Persona	Cheque de pago
Empleado		Mandatorio	Opcional
Persona	Opcional		
Cheque de pago	Mandatorio		

Tabla 1.5. Relación entre empleado, persona y cheque de pago.

Persona
Empleado Opcional
Empleado
Persona Mandatorio
Cheque Opcional
Cheque
Empleado Opcional

1.3.1.3.1 INTEGRIDAD ESTRUCTURAL CONDICIONAL DE LOS DATOS

La integridad estructural condicional especifica la cardinalidad para las relaciones entre los datos. La cardinalidad es el número de ocurrencias permitidas a un dato de cada uno de los lados de la relación. Cuando existe una arquitectura común de datos, la cardinalidad es documentada por medio de la integridad de los datos, y no con la ayuda de la estructura de los datos. La cardinalidad se documenta por medio de una tabla de reglas de integridad estructural condicional.

Ejemplo 1.10:
Un estudiante, puede tener cero, uno o más grados universitarios. Un estudiante no graduado no puede tener ningún grado. Los estudiantes graduados, a cambio si pueden tener uno o más grados. (Ver tabla 1.6.)

Tabla 1.6. Cardinalidad de una relación.

Estudiante		
No graduado	0	Grado Preventivo
Graduados	1,M	Grado Mandatorio

La tabla de integridad estructural condicional muestra la cardinalidad. En un modelo relacional, esto es parecido a las relaciones entre las diferentes entidades, que pueden ser uno a uno, uno a muchos y muchos a muchos. Otra forma de definición es a través de la notación formal:
Estudiante = { No graduado | Licenciado | Postgrado | Doctorado }
Cualquiera de esas formas es aceptable para definir tanto la integridad estructural de un dato, como la integridad estructural condicional. El objetivo final es definir de manera apropiada y consistente la integridad estructural y hacerla entendible y aceptable para el usuario y los clientes en general.

1.3.1.3.2 INTEGRIDAD ESTRUCTURAL REFERENCIAL

La integridad referencial es la parte de la integridad estructural de los datos que asegura la existencia de parentesco entre datos para cada ocurrencia existente subordinada. Una ocurrencia subordinada de un dato no puede ser añadida si no existe una ocurrencia "padre" ya existente, y por otro lado una ocurrencia " padre" de un dato no puede ser borrada, si alguna ocurrencia subordinada sigue existiendo. Por ejemplo el pedido de un cliente no puede ser aceptado si no se tiene el registro de su cliente correspondiente y un cliente no puede ser borrado si tiene pedido existente. La integridad referencial generalmente se usa dentro de la arquitectura común de datos y no se necesita detallar minuciosamente cada relación entre los datos existentes.

1.3.1.4 INTEGRIDAD DE LA RETENCION (CONSERVACION) DE LOS DATOS

Los datos siempre están en movimiento y sufren cambios constantes. En muchas ocasiones son descartados en base a alguna operación actual dentro de la empresa, sin que se consideren sus valores futuros, para alguna aplicación mas adelante. La integridad de la retención de los datos es otro subconjunto de la integridad de los datos, que crea y

especifica el criterio necesario, para prevenir la pérdida de datos cruciales o de suma importancia en el proceso de actualización o en el caso de que sea borrados algunas partes de la información. La retención debe considerar los valores futuros de estos datos y debe de determinar, que dato tiene que ser retenido y de que manera lo hará. La retención debe de ser capaz de considerar los valores futuros, para determinar una necesidad desconocida o escondida de un dato. La integridad de la retención es especificada, para las ocurrencias y las características de los datos. Las reglas de la retención de ocurrencias de datos especifican, para que periodo de tiempo una ocurrencia debe de ser retenida y que es conveniente hacer antes de que sea borrada. También se aclaran los procedimientos, que preservan la importancia histórica de una ocurrencia de datos con sus respectivos valores característicos. Los posibles procedimientos, para preservar el valor histórico y la significancia de las ocurrencias de los datos, consisten en la creación de auditorias, para una parte o la totalidad de las características de los datos, dentro de cada ocurrencia de un dato. Para hacer esto es necesario mover las ocurrencias dentro de un archivo histórico, con el propósito de realimentar los valores de los datos, por medio de decisiones que permiten que los datos alcancen un mayor nivel de comprensión, y con esta base tomar la decisión de que ocurrencias van a ser borradas y cuales serán preservadas. Podemos tomar como ejemplo la retención de una ocurrencia común como es la situación cuando un empleado abandona una empresa. Se puede tomar como regla que la ocurrencia de este empleado sea borrada después del treinta y uno de enero del año siguiente de su ausencia de la empresa. Todos los valores característicos de esta ocurrencia se deben de mover a un archivo histórico del empleado en cuestión. Con esto nos podemos asegurar, que si por alguna razón necesitamos este archivo, la información guardada podrá fácilmente ser usada, para un propósito específico.
Las reglas para retener a las características de los datos especifican los procedimientos necesarios cuando un valor de un dato característico es actualizado o borrado. Esto es importante, para preservar la significancia histórica de este valor característico. Al igual que en el caso pasado son necesarias auditorias, para seguir los cambios en los valores de los datos y de esta manera aumentar el nivel de la calidad de estos datos.

Ejemplo 1.11:

Empleado
La fecha de nacimiento puede ser actualizada sin que se guarden los valores anteriores. Nunca se debe de borrar. La edad del empleado puede ser actualizada o borrada en cualquier momento sin guardar el valor histórico. El nombre del empleado puede ser actualizado, y el valor anterior se debe de grabar en un archivo de auditoria. Nunca se debe de borrar.

Las reglas de la retención de datos son mantenidas dentro de las características de los mismos, de la misma manera como las otras reglas de integridad de los datos. Las reglas sobre las ocurrencias de un dato son mantenidas junto al sujeto y el significado del dato. Las reglas de retención de las características de los datos son mantenidas dentro de las mismas características. Lo más importante es que las reglas se definan de tal manera que puedan implementarse.

1.3.1.4.1 INTEGRIDAD DE LOS DATOS DERIVADOS

Un recurso para datos contiene una cantidad considerable de datos derivados. Tanto los procedimientos de derivado como los procedimientos, de mantenimiento de los datos derivados necesitan una documentación apropiada, para asegurar una alta calidad de los datos. Los datos importantes junto con sus replicas son de extrema importancia dentro la organización de recursos de datos. La existencia de datos redundantes, sus replicas y su mantenimiento también necesita ser documentadas para elevar el nivel de calidad.

La derivación de datos es el proceso de crear un valor de dato a partir de uno o más existentes valores de datos, por medio de un algoritmo de derivación. Un dato derivado es cualquier dato creado por medio de un proceso de derivación. Un dato derivado activo, es cualquier dato que se deriva a partir de otro dato o datos, con sus valores existentes en este momento y que puedan ser sujetos a cambios. Los datos derivados activos tienen que ser rederivados siempre y cuando los valores de los datos originales cambien o cuando aparece algún otro valor de dato contribuyente. El dato derivado estático es cualquier dato que se origina a partir de otros datos contribuyentes cuyas características ya no existen o si sus valores nunca cambian y son constantes. Prácticamente nunca existe la necesidad de rederivar un dato derivado estático.

La integridad de los datos derivados es un subconjunto de la integridad, que especifica los criterios para la derivación de los datos y su mantenimiento. Estas especificaciones incluyen las características de los datos y los procedimientos de derivación. Un diagrama para la derivación especifica las características de los datos contribuyentes y al dato derivado. Los elementos del diagrama constan de rectángulos, que representan las características del dato, y las flechas representan la dirección de la derivación. El nombre del dato característico se coloca dentro del rectángulo. La dirección de las flechas debe de ir desde los datos contribuyentes hacia el dato derivado. El proceso de derivación de un dato viene implícito en las flechas. El problema con la mayoría de los datos derivados, es que una vez derivados ya no se actualizan, mientras que los valores de los datos contribuyentes cambian o hay nuevos datos contribuyentes añadidos. Por esto es importante el proceso de rederivación, que valida los valores de los datos derivados. A este proceso se le llama mantenimiento de los datos derivados. La cantidad de veces que se realiza este proceso durante un lapso de tiempo se determina por el criterio de la persona encargada, puede ser diario, a media noche, mensual o anual.

En particular los datos derivados activos necesitan guardar la versión del cambio de valor. La identificación de la versión se vuelve extremadamente importante cuando comparamos para analizar tendencias o hacer proyecciones.

1.3.1.4.2 INTEGRIDAD DE LOS DATOS REDUNDANTES

Los recursos de datos existentes a menudo tienen redundancias, aun que esta se desea limitar únicamente a las llaves primarias o secundarias. La redundancia de un dato se da cuando las características de este se repiten en uno o mas lugares. En muchas ocasiones los datos redundantes fueron creados, almacenados y mantenidos independientemente uno del otro y casi siempre la empresa no los conoce. Por esta razón cuando los valores se actualizan inconscientemente casi siempre no coinciden. Este tipo de datos debe ser

identificado, documentado y mantenido adecuadamente, para asegurar la calidad de la información. Por medio de un proceso de mantenimiento se pretende llegar a que cada dato redundante tenga un valor consistente. Para una mayor representación se usan diagramas de datos redundantes. Los componentes de estos diagramas constan de rectángulos y flechas dirigidas. Los primeros representan los datos característicos y las flechas las relaciones con los datos redundantes.

Ejemplo 1.12:
La fecha de nacimiento de un empleado puede existir en el archivo de empleados, en la nómina y en el archivo de capacitación. El archivo de empleados es designado como el recurso oficial de datos, mientras que los otros dos archivos son recursos redundantes que tienen que ser actualizados desde el recurso oficial. Cada vez que alguna fecha de nacimiento es agregada o cambiada en el archivo de empleados, es inmediatamente agregada o cambiada en la nómina y el archivo de capacitación.

1.3.1.4.3 INTEGRIDAD DE LA REPLICAS DE LOS DATOS

Las replicas de los datos sirven, para que ellos sean distribuidos a varios sitios para su uso posterior. La replicación de datos es un proceso que crea copias exactas de datos existentes en un archivo, para pasarlos a otro archivo. La réplica de los datos no es lo mismo que los datos redundantes. La diferencia radica que cada réplica es planeada, conocida, documentada y apropiadamente mantenida.

1.3.2. PRECISION DE LOS DATOS

El segundo componente de la calidad de los datos es su exactitud. La exactitud trata sobre la forma de cómo determinar si los datos que tenemos almacenados en los recursos informáticos, representen de manera más exacta el mundo real. Cuando hablamos de calidad de datos los términos exactitud y precisión se consideran sinónimos. Al principio se tiene que identificar el nivel actual de precisión y también los ajustes necesarios para alcanzar las necesidades de la empresa. Esto asegura que el nivel actual ya se conoce y por lo tanto si es posible también alcanzar el nivel deseado. Hay que tener en cuenta que no existe algún nivel de exactitud predefinido. Cada organización o empresa determina el nivel deseado de precisión sobre las características de sus datos con el fin de cumplir con sus necesidades. Cuando el nivel existente de exactitud es determinado para cada dato, hasta para los que toman valores muy variados, los ajustes llevaran a cabo la tarea de subir al nivel deseado. Esto es posible siempre y cuando exista una arquitectura común de los datos. La precisión o exactitud incluye varios detalles a tratar, tanto directos como indirectos. Las cuestiones directas incluyen los métodos usados para identificar objetos y eventos del mundo real, los métodos de reunificación de datos sobre estos objetos y eventos, el tiempo de colección de datos y las metodologías usadas para captura de los datos dentro de los recursos informáticos. Los conceptos directos abarcan la precisión, escala, resolución, coherencia, el nivel de detalle, el grado de redundancia, instancias de los datos y volatilidad, linaje, versiones y en general todo aquello sobre la manera en que el dato representa el mundo real.

Los puntos indirectos incluyen la persona que captura e introduce los datos, la que declara que el valor de un dato sea correcto y las personas internas de confianza que manejan los

datos La confianza en las personas u organizaciones que colectan los datos o los métodos de captura se reflejan en la confianza en los datos, y en su nivel de precisión. Falta de confianza en algunas piezas de datos puede destruir la confianza en un recurso de datos integrados.

1 3.2.1. ACTUALIDAD DE LOS DATOS

La actualidad es una medida que permite determinar qué tan relevantes son los datos en comparación con el mundo real o también el grado de caducidad que tienen. La actualidad depende de la volatilidad, de las instancias y de la colección de frecuencia de los datos.
La instancia es el punto en el tiempo o el periodo para el cual los valores de los datos representan de manera exacta el mundo real. La instancia debe de ser conocida para cada dato. Por ejemplo el empleo de un cliente puede cambiar varias veces al año, pero la fecha de nacimiento es permanente. La volatilidad es la manera de qué tan rápido los datos que representen el mundo real se vuelven obsoletos o inadecuados, y esto depende de qué tan cambiante sea el mundo real. La volatilidad es la frecuencia con la que los datos dejan de representar el mundo real. La volatilidad se debe de conocer para definir la actualidad de los datos. Por ejemplo los signos vitales de un paciente durante la operación de su corazón cambian cada segundo, pero el nombre de la persona permanecerá el mismo durante toda su vida. La frecuencia de recolección es la frecuencia en la cual los datos son recolectados a partir del mundo real y debe de ser conocida como cualquier dato. Por ejemplo los signos vitales del paciente durante la operación deben de ser recolectados cada segundo, en cambio la población de una ciudad se recolecta cada diez años. Existe una relación muy importante entre la frecuencia de recolección y la volatilidad de los datos. La frecuencia de recolección debe de encajar con la volatilidad. Los datos que no se mantienen frescos se deterioran y esto nos lleva a la falta de actualidad de los recursos informáticos. Por ejemplo si los signos vitales del paciente durante una operación son tomados cada hora, el paciente puede morir antes de que los cambios notifiquen al equipo médico y se pueda reaccionar correctamente. Por otro lado mediendo la población de la ciudad cada día es inútil, ya que no va a mostrar cambios significativos.

La actualidad de los datos es mantenida ajustando la frecuencia de recolección de datos a la frecuencia en que un evento ocurre o un objeto cambia en el mundo real. Si los cambios ocurren más frecuente que la captura de los datos, algunos eventos se perderán y los datos no van a ser actuales. Los eventos de cambios pueden ser continuos o discretos. Si los eventos son continuos, como un flujo, se escoge un intervalo arbitrario, para capturarlo en la frecuencia deseada por la empresa, puede ser cada hora o cada día. Las diferentes empresas tienen diferentes niveles de actualidad, el intervalo debe de tener la mayor frecuencia de recolección deseada. Si un evento es discreto como por ejemplo un accidente automovilístico, la recolección de frecuencia de los datos debe de ser durante el evento o lo más pronto después de que este ocurra, según la necesidad de la empresa. Si diferentes empresas tienen diferente frecuencias de recolección de los datos, la empresa que tenga la recolección más pronta, proporcionara la mejor información. El tiempo de captura es la rapidez con la que los datos son capturados, introducidos dentro de los recursos informáticos, distribuidos a los diferentes sitios y disponibles para ser usados. Es la medida que determina qué tan rápido los recursos representan los cambios de la información. El tiempo de captura incluye la recolección de los datos de la manera más eficiente, la edición

y la corrección con el menor retraso posible. La actualidad de los datos es mejorada achicando el lapso de tiempo entre la recolección y la disponibilidad para el usuario de estos datos.

1.3.2.2. LINAJE Y HERENCIA

El origen de los datos es el lugar de donde provienen. Dicho de manera más específica es donde los datos se colectan, crean, se miden, generan, derivan, modifican o agregan. Este lugar puede ser un campo de medición, un laboratorio de análisis o una encuesta. El origen de los datos es un sitio específico para el almacenamiento de datos y de donde estos se pueden obtener. Un origen primario de datos es el primer lugar donde los datos se almacenan después de su creación. Un origen de datos secundarios puede ser cualquiera donde los datos son adquiridos desde otro sitio y están almacenados sin cambios o modificaciones. Los datos en un origen secundario pueden provenir tanto de un origen primario como de otro secundario. Si por alguna razón estos datos son cambiados o modificados de alguna forma en este sitio, entonces este sitio se convierte en un origen primario, para estos nuevos datos. Un origen oficial de datos es cualquier sitio, donde los datos oficiales y las grabaciones de referencia son almacenados. Cuando existe redundancia en los datos, se debe de tomar el origen oficial para desarrollo de algún recurso integrado o para crear réplicas en otros sitios.

Un origen no oficial de datos es cualquier sitio, donde los datos no tienen sus grabaciones de referencia en el mismo lugar. Dicho de otra manera es un sitio donde se localizan datos redundantes y no se van a usar para la creación de otro recurso integrado de datos. Un recurso oficial puede ser diferente de un recurso primario de datos.

Para la aclaración de este punto se puede tomar el ejemplo cuando los datos se originan en un campo de medición, que en realidad es su origen y después son almacenados en la base de datos principal de la empresa, que es su fuente primaria. Otro ejemplo es cuando varias organizaciones transfieren los datos de un campo de medición hacia alguna parte central común sin alteración, que se vuelve un origen secundario de datos. Esta localidad central es en realidad un origen oficial de datos para estas mediciones en el campo. El seguimiento de los datos (data tracking) documenta el movimiento de los datos desde su origen, pasando por los sitios primarios, hasta llegar a los secundarios. Este seguimiento documenta cualquier alteración o modificación, así como el flujo de nuevos datos y la creación de datos derivados o agregados. Es un proceso que nos ayuda a comprender el manejo y el movimiento de datos dentro y fuera de la organización. El seguimiento es un proceso importante, para determinar el linaje de un dato. El linaje es comunmente usado para encontrar descendientes biológicos culturales, pero puede ser aplicado a los datos. El linaje de los datos es un proceso que sigue el camino de los datos, desde sus valores originales, hasta sus lugares y situaciones actuales. Esto incluye determinar donde se originan los valores de los datos, donde se almacenan, y de que manera son alterados o modificados. Es en realidad la historia de los datos, su origen, las modificaciones que sufren hasta llegar a su forma y ubicación actual. El linaje sirve para documentar las características específicas de un dato o un conjunto de datos, para alguno o varios sucesos de datos y también para algún archivo o base de datos. El linaje de los datos casi siempre es documentado por medio de un diagrama de linaje, donde se especifica el flujo y las características del conjunto de datos que nos interesa con la ayuda de comentarios.

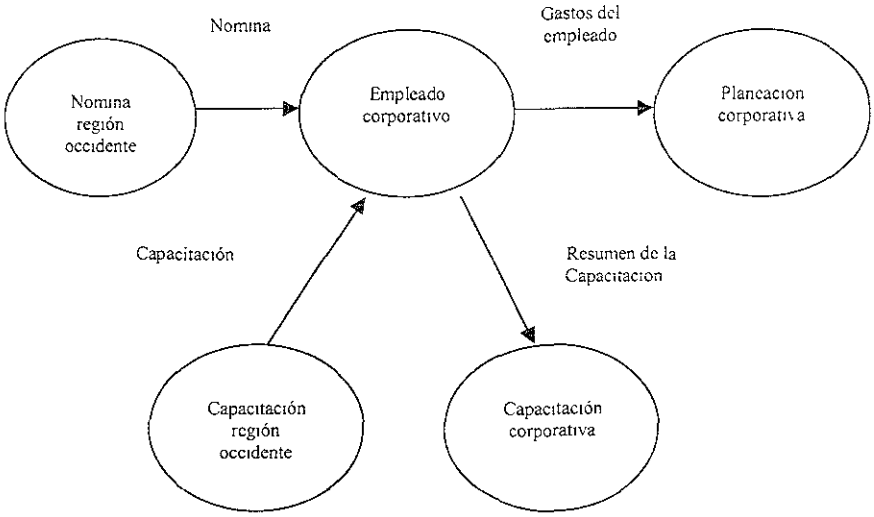

Figura 1.2 Diagrama de linaje de datos.

La figura 1.2 representa un diagrama de linaje de datos y muestra el movimiento de los conjuntos de datos desde su origen hasta su ubicación actual. En el siguiente diagrama los sitios de los datos son indicados con círculos ovalados. Ellos representan el almacenamiento físico de los datos con el nombre de cada sitio escrito adentro del círculo. Los conjuntos de datos y su movimiento entre los diferentes sitios se muestran por medio de las flechas. Como ejemplo se toman los datos de la nómina y la capacitación de los empleados. Los datos de la nómina se mueven de la región occidente hacia los datos corporativos de los empleados. Los datos de capacitación se mueven de la capacitación región occidente hacia los datos corporativos. Los gastos del empleado se mueven desde los datos corporativos hacia los datos de planeación corporativa. El resumen de la capacitación se mueven desde los datos corporativos de los empleados hacia los datos de capacitación corporativa.
Dentro de los diagramas de linaje aveces se usan enunciados de apoyo que describen la historia y el ciclo de vida de los datos. Estos diagramas también pueden contener procedimientos para derivación de datos.

Por otro lado el linaje de los datos no es lo mismo que los datos derivados o datos distribuidos. El linaje describe la historia que representa el ciclo de vida desde el origen de los datos. Los datos derivados describen las características continuas y los procedimientos para derivar los valores de los mismos. La distribución de los datos describe las replicas y la distribución misma desde el origen oficial, hacia los orígenes no oficiales de los datos.

La herencia es algo transmitido, adquirido de algún predecesor, o algo que ya se tiene como resultado de un proceso natural de creación. La herencia al igual que el linaje

generalmente se usa en la biología para referirse a los descendientes, pero puede ser aplicada también a los datos. El significado de la herencia de los datos se refiere al lapso de tiempo desde que se originan siguiendo su movimiento hasta el sitio actual que tienen asignado. También incluyen una información que contiene las alteraciones y cambios que sufren desde ese proceso, el linaje de los datos describe los pasos que los datos siguen y de que forma son transmitidos durante su ciclo de vida. En cambio la herencia nos ayuda a entender el contenido original, su significado y como este significado cambia durante su movimiento de un sitio a otro.

1.3.2.3. DATOS TEMPORALES

Los datos temporales se han vuelto cada vez más importantes tomando en cuenta la siempre mayor complejidad de los recursos de datos y también de los procesos analíticos. Temporal, significa algo relacionado y limitado por el tiempo, proviene del verbo latín tempos, que significa tiempo. Los datos temporales son todos aquellos que representen un punto o intervalo de tiempo teniéndolo como su componente principal. Los datos temporales son un componente importante de la actualidad, existen varios diferentes tipos de características de los datos temporales que son:

La fecha del evento es el lapso durante el cual el dato tiene significado en el mundo real.

La fecha de identificación es el momento en que la empresa descubre dicho evento

La fecha de colección, durante el cual se reúne la mayor cantidad de datos sobre este evento

La fecha de entrada cuando alguna transacción introduce este dato dentro de los recursos informáticos.

La fecha de distribución, que ocurre cuando el dato es distribuido en un lugar en especial.

Los datos temporales deben de ser reunidos al nivel de detalle requerido por la empresa y de acuerdo a sus necesidades. Por ejemplo los tiempos geológicos como el Mioceno y el Jurásico son usados para determinar la edad de las rocas. Siglo, año, mes y día son usados para la fecha de nacimiento. Horas, minutos, y segundos son usados para pacientes en cirugía y las pequeñas fracciones de segundos se usan para los experimentos de la física nuclear.

En un punto o intervalo de tiempo se colectan las ocurrencias de características de los datos. Un punto en el tiempo es una fracción única para una ocurrencia que representa alguna característica de un dato. Un intervalo de tiempo tiene su inicio y final para cada ocurrencia de la característica de un dato. El tipo de un dato temporal se puede identificar con una palabra común como es evento, identificación, colección, inicio y final. Es importante mencionar que por experiencia las bases de datos temporales manejan los datos temporales mejor que las bases de datos relacionales. Una base da datos relacional es orientada para los valores actuales y no puede manejar adecuadamente los datos temporales. Los datos temporales como son los de auditoria de pruebas, datos históricos y archivos de datos, pueden ser mantenidos en una base relacional, pero este tipo de base de datos no tiene la capacidad creativa, para regenerar los valores pasados o futuros de estos datos. Las bases de datos temporales, conocidas como bases de datos relacionales de tiempo, tienen la capacidad para almacenar datos temporales y manejar datos que se

originan a raíz de nuestros datos temporales. Esto le proporciona la capacidad de crear valores validos pasados o futuros basados en los valores actuales de datos temporales.

1.3.2.4. VERSIONES DE LOS DATOS

Las versiones de los datos representan el componente más crítico de la actualidad, como característica. La versión de los datos es un conjunto de valores de datos que representa el mundo real en un punto específico del tiempo. Los datos temporales son un componente clave para los nombres de los diferentes tipos de versiones de un dato y su definición, porque ellos muestran el punto o intervalo en el tiempo para el cual estos valores de datos son válidos. La versión de los datos es crucial para determinar el tiempo y el método que se va a usar para preparar nuestra información. El método de preparación es un componente clave para las versiones de datos, como es una búsqueda usada para crear un conjunto de valores. El método usado en la preparación de los datos debe de ser documentado, para entender las versiones de los datos y como han sido desarrollados estos valores. La comparación de diferentes conjuntos de análisis o sumarización, que se basan en diferentes métodos, para la preparación de los datos no tienen significado sin las versiones de los mismos, estas versiones generalmente tienen que ser conocidas, para que los datos puedan cumplir con su misión de manera apropiada.

1.3.2.5. MULTIPLES MODIFICACIONES DEL ORIGEN DE LOS DATOS

Esta situación se presenta cuando varias empresas usan una única fuente de información y hacen simultáneamente ajustes para sus necesidades. Como los ajustes son hechos de manera independiente y no están sincronizados de ninguna manera como resultado se presentan múltiples orígenes de los datos, que entran en conflicto con las versiones de datos. La calidad de la información puede ser sumamente afectada, si no existe sincronización apropiada entre estos orígenes. Existen dos técnicas, para resolver este problema, dado que las versiones de datos no van a coincidir. La primera técnica trata de establecer una fuente única de datos, que continúe las versiones mas recientes de estos datos, basándose en su origen. Los datos actualizados siempre van a ser disponibles en este recurso. Las empresas que proveen los datos son la fuente primaria, pero no son el origen oficial de los datos. El origen central es la fuente oficial de los datos, aun siendo un origen secundario.

La segunda técnica establece nombres únicos de los datos, para cada recurso primario y desarrolla las versiones de nombre de los datos usando los nombres provenientes del recurso primario. Los datos son actualizados y se mejoran en sitios descentralizados, los cuales tienen las dos funciones, la de origen primario y la de origen oficial.

1.3.2.6. ACTUALIZACIONES PROACTIVAS Y RETROACTIVAS

Las bases de datos temporales permiten tanto actualizaciones proactivas como retroactivas. Las actualizaciones proactivas tienen una característica en particular, que permite que una actualización hecha en alguna fecha actual, se haga efectiva en una fecha futura. Este tipo de actualizaciones requieren de una fecha de introducción y una fecha efectiva. En el caso

de un banco se puede definir un cambio de la taza de interés a partir de una fecha futura. El dato se actualiza hoy, pero entra en vigor según lo indica la fecha de inicio.
Este tipo de actualizaciones es benéfico, porque pueden marcar el paso a sucesos, que ocurran en el futuro, y empezar una serie de otras actualizaciones. Hay que tener en cuenta, que si los datos no se usan apropiadamente con respecto al tiempo, no se podrá obtener algún beneficio de ellos, por esto es necesario que se manejen de manera correcta. Las actualizaciones retroactivas son las que permiten, hacer una actualización de los recursos de datos hoy, y que la misma fue ya efectuada en alguna fecha del pasado. También requieren de una fecha de entrada y otra efectiva. Por ejemplo cuando en alguna cuenta de un cliente no se le deposita de manera apropiada el dinero, esto permite que se cometa otro error, porque se le cobran al cliente más intereses y cargos, cuando en realidad la culpa es del banco. La corrección del problema se hace por medio de una actualización retroactiva, que elimina los cargos coaccionados por el error. Las actualizaciones retroactivas son buenas, porque hacen los datos efectivos, en la fecha correspondiente con un evento actual del mundo real. Uno de los problemas que se puede presentar con este tipo de actualización, es su capacidad de cambiar los datos históricos y cualquier análisis posterior, que se base sobre estos datos nos va a llevar a cometer errores. La mejor manera, para manejar las actualizaciones retroactivas es documentando los datos actualizadas, como una nueva versión de estos datos. Cualquier derivación o análisis de los datos actualizados va a contener la versión anterior del dato, antes de la actualización retroactiva. Cuando aparece una versión más reciente de dato, esto puede indicar que cualquier derivación previa o análisis puede ser un error.

1.3.3. COMPLETES DE LOS DATOS

El tercer componente de la calidad de los datos es la completes. La completes es el indicador más importante, que se encarga de determinar si los datos pueden cumplir con la demanda actual y futura de información dentro de los recursos informáticos. De la misma manera muestra la cantidad de datos disponibles para su uso dentro de la empresa. Determina que datos son necesarios para cumplir con la demanda de información, y asegura la captura y mantenimiento de estos datos para su uso efectivo en el momento en que se requieren. La completes se lleva a cabo dentro de los recursos de datos, por medio de inspecciones e inventarios de los recursos. La inspección de los recursos de datos determina a un nivel muy alto las necesidades de datos de la empresa y cuales de estos datos son disponibles. Para cumplir con esta tarea se basa en esquemas que clasifican los datos a un nivel suficiente. El primer paso casi siempre es determinar las prioridades que permitan un inventario detallado de los recursos de datos. Este inventario consta en determinar la demanda de datos de la empresa y su disponibilidad, basándose en sus características y cualidades. Antes de esto, la inspección ya debería haber determinado las necesidades y los detalles mas mínimos acerca de los datos. Usados juntos estos dos procesos siempre descubren los recursos escondidos de datos disponibles de manera inmediata para su uso, dentro de la empresa.
La inspección de los recursos consta de tres partes: la inspección de necesidades, la inspección de disponibilidad, y el análisis de la inspección. Todo esto se hace con el único propósito de cumplir con la demanda de información. En la última parte el análisis compara las necesidades con los datos disponibles, para así poder determinar que datos ya existen y cuales necesitan ser adquiridos. La prioridad de las estrategias de los negocios, es conseguir

los datos necesarios, para poder identificar datos cruciales requeridos, para que en caso de que no existan poder planear como adquirirlos y satisfacer las necesidades inmediatas a largo plazo.
El primer paso determina las necesidades conociendo las características de los datos necesarias para las actividades empresariales. En el segundo paso, el inventario de disponibilidad identifica la existencia y características de los datos obteniendo una información todavía mas detallada sobre ellos. En la fase final, el análisis permite comparar los resultados de los primeros dos pasos y así saber con precisión de que datos disponemos y cuales se deben de adquirir. La conveniencia de los datos es una parte de la completes, que indica qué tan convenientes son los datos para una tarea o actividad especifica. Los mismos datos pueden ser más convenientes en una actividad y menos convenientes en otra. La razón de los tres pasos anteriores es precisamente determinar este aspecto clave de conveniencia.

1.3.3.1 CALIDAD DE LOS METADATOS

En la mayoría de las empresas los valores de los metadatos son muy disparejos y carecen de integridad. Si es que existen, son incompletos, mal escritos, no actualizados, difíciles de encontrar y entender. Muchos de los metadatos son mantenidos por empleados capacitados que se retiran o cambian de trabajo, lo que resulta en la completa perdida de estos metadatos. Las dificultades aumentan, proporcional al tamaño y a la complejidad de los recursos como, distribuciones de los datos, sistemas de data warehouse, y sistemas geográficos de información. Paralelamente a esto cada vez mas personas recolectan y almacenas volúmenes de datos mayores. Una empresa necesita de desarrollar un conjunto suficientemente grande de metadatos que puedan describir de manera adecuada los recursos de datos y situarlos en un orden formal, para que pueda responder a la demanda de información. Los metadatos de alta calidad, cuando son disponibles proporcionan un apoyo a los manejadores de recursos de datos y a los usuarios de la información en el entendimiento de los recursos de datos. Sólo de esta manera estos recursos podrán ayudar de una manera integra en las actividades del negocio.

En la mayoría de las organizaciones el estado de los metadatos es malo. Sus valores son tan ilógicos como lo son los de los datos. Esto a su vez pone barreras a la integración y intercambio de datos entre empresas que no cumplirán con la demanda de información del negocio. El estado actual de los metadatos tiene que ser entendido, para que puedan ser mejorados y ser plenamente útiles en el recurso integrado de los datos.

Los metadatos son siempre dispersados en documentos diferentes y son difíciles de localizar. Por otro lado cuando son encontrados, ellos no siempre explican de manera completa y minuciosa el contenido, el significado y la calidad de los datos. Cuando se tienen metadatos en gran cantidad y además muy disparejos, como es el caso cuando están repartidas entre varias jurisdicciones y áreas, estos metadatos son muy difíciles de integrar. Esto impide que los trabajadores entiendan los datos que tienen a su disposición. En algunas situaciones, los metadatos en grandes cantidades y no integrados, son mas difíciles de integrar que los datos mismos. Los trabajadores capacitados, los clientes de negocio y programadores, retienen una cantidad considerable de metadatos que nunca se llega a documentar. Esto es una permanente perdida de información en los recursos de datos.

En realidad las organizaciones no hacen ningún intento de mejorar sus metadatos. Las empresas que dedican parte de su esfuerzo tienen poco progreso. En el mismo tiempo diferentes personas continúan capturando y creando nuevos datos, por medio de diferentes métodos. Estos datos no son documentados mejor que los datos que ya se tienen, a pesar de la existencia de mejores herramientas y técnicas para documentación. De hecho la situación se pone peor aun, porque son documentados de manera diferente, por medio de diferentes herramientas lo que resulta en la creacion de mas datos sin calidad. Esto crea dos problemas principales. El primero es la falta de integridad, consistencia y completes que provoca la situación ya mencionada. La segunda es que dado esta situación, los conjuntos de metadatos separados no pueden proporcionar una vista comprensiva de los datos disponibles en la empresa. Esto crea un gran dilema para las empresas dado que, se tiene que escoger entre tener buenos metadatos, para poder utilizar por completo el recurso integral de los datos o continuar de desarrollar datos que cumplan con la demanda de información actual.

Manejar los metadatos es por lo menos igual de difícil que manejar datos. En el caso ideal los metadatos representan los datos de una manera precisa de la misma manera que los datos deben de representar al mundo real. Generalmente ninguna de estas dos situaciones se cumplen. La calidad de los metadatos depende enteramente de la integridad de los datos que las personas crean. Un verdadero principio para los metadatos es que los especialistas tienen que reunir y documentar los datos de la mejor manera posible, para que puedan servir al cliente. El mayor problema con los metadatos es cuando se desconozcan sus valores. Los metadatos a veces no existen, no están disponibles, y no son comprensivos. La mejor aproximación, para desarrollar los metadatos es documentar los productos originales de datos y añadir mas información que se crea con el paso del tiempo. El proceso de creación de los metadatos es un proceso de investigación. Una vez creado el repositorio de los metadatos se debe de crear una referencia cruzada, por aplicativo o por sistema. Por medio de adiciones de los metadatos, ajustes en las referencias cruzadas y mejorando los metadatos comunes se podrá entender el nivel de calidad y se podrá mejorar este nivel. Los expertos de negocio, y los expertos de sistemas deben de trabajar en equipo, para así poder documentar los metadatos. Estos especialistas son los únicos que tienen el conocimiento y el perfil, para cumplir con la tarea. El desarrollo de los metadatos comunes tiene que ser prioridad de mas alto nivel, sea para un proyecto de producción o para desarrollo de nuevos datos.

1.3.4. MANEJANDO LA CALIDAD DE LOS DATOS

Para muchas empresas la calidad de los datos es un gran problema. Esto se debe a que en general no se toma mucho en cuenta. La mayoría de los datos nunca han recibido auditorías o validación de algún modo. Mejorar la calidad de los datos es un proceso complejo que empieza determinando la existente calidad de datos y acaba cuando se alcanza un consistente nivel de la calidad de la información.

1.3.4.1 MEJORANDO LA CALIDAD DE LOS DATOS

Existen varios diferentes nombres que pueden ser usados para nombrar este proceso, algunos de ellos son: limpieza de datos, purificación de datos, rastreo etc. Todos estos nombres son de alguna manera sinónimos con el mejoramiento de la calidad. El mejoramiento se puede hacer de manera prospectiva o retrospectiva. Existen dos tipos básicos de mejora. Mejoras prospectivas es el nombre del proceso que se encarga de mejorar la calidad solamente de los datos nuevos, que entran en los recursos de datos, pero generalmente ignoran mejorías en la calidad de los datos que ya existen en los recursos informáticos Mejoras retrospectivas es el proceso de mejorar la calidad de los datos ya existentes. Casi siempre se empieza con el tipo de limpieza prospectivo de datos y después se pasa al proceso retrospectivo, lo que permite bajar los costos y aumentar la rapidez de la tarea asignada.

1.3.4.2 CRITERIOS DE LA CALIDAD DE LOS DATOS

Los procesos para mejorar la calidad de los datos empiezan estableciendo los criterios a seguir. Estos incluyen todo lo mencionado antes sobre la integridad, precisión y completes. Los criterios de la calidad de datos existentes son los criterios que son documentados en los recursos actuales existentes. Los criterios deseados de calidad son las reglas a las que se quiere llegar para poder soportar una mayor demanda de información. Ambos procesos no pueden ser completamente automatizados, porque requiere de personas capacitadas que conozcan los problemas de la empresa y de calidad. Normalmente los criterios de calidad existentes son conocidos por pocas personas y casi nunca documentados. Las herramientas automatizadas existentes pueden ayudar a las personas que conocen el proceso de documentar sus análisis, pero no pueden reemplazar el proceso de descubrimiento de los criterios existentes.

1 3.4.3 TECNICAS DE LA CALIDAD DE LOS DATOS

Existen dos técnicas básicas para mejorar la calidad de los datos existentes. La técnica inductiva es un proceso que va desde lo particular a lo general, donde los datos existentes son analizados y los criterios son desarrollados a partir de estos datos. Se evalúa si los criterios viejos son aceptables en su forma original, en caso de que no, entonces se apliquen los ajustes necesarios. La técnica deductiva es un proceso que va de lo general a lo particular. Primero se desarrollan y aplican los criterios de calidad necesarios, en lugar de determinar el nivel actual de rendimiento o tratar de asegurar que los nuevos datos capturados sean correctos. Los datos de baja calidad existentes se ajustan o modifican si es necesario, según los criterios desarrollados, para alcanzar el nivel deseado. Tanto la técnica inductiva como la deductiva pueden ser combinadas dentro de un proceso cíclico para obtener un resultado optimo. Esto permite el desarrollo de los criterios existentes, con la ayuda de la técnica inductiva. Después de analizarlos y mejorarlos hay que demostrar que soportan la demanda de información. Entonces los nuevos criterios se aplican por medio de la técnica deductiva sobre todos los datos, nuevos y viejos. El proceso de calidad de datos documenta y mejora la calidad usando técnicas deductivas y inductivas. Es un proceso sistemático que examina los recursos de datos y posteriormente se ajusta a las necesidades

de la empresa para alcanzar el nivel deseado de calidad. Se necesitan auditorias constantes y rutinarias sobre los datos para evaluar su integridad y exactitud.

1.3.4.4 PROCESOS PARA ASEGURAR LA CALIDAD

El proceso de calidad incluye cinco pasos:

Descubriendo el problema de la calidad de datos
Es relativamente fácil para una empresa determinar si la calidad de los datos no esta al nivel deseado y tomar una decisión sobre que ajustes se deben de hacer. En la mayoría de las situaciones las empresas ya tienen planeado este proceso.

Entendiendo la calidad de datos existente
La tarea difícil es entender y documentar la calidad existente. La documentación se logra por medio del proceso inductivo. Una pequeña parte de los recursos informáticos es seleccionada, de preferencia alguna parte clave, para las operaciones de los negocios. Sobre la información seleccionada se efectúa un análisis. Los datos se revisan y documentan usando las mismas técnicas que sirven para definir su integridad y su precisión. El resultado es un conjunto de criterios existentes de calidad de datos que representan la porción de información seleccionada.

Determinando el nivel de calidad deseado
El nivel deseado es el que soporta la demanda de información. Se alcanza por medio de revisiones y modificaciones de los criterios existentes sobre la calidad de datos. Generalmente los ajustes constan en un aumento de los criterios, pero en algunas situaciones puede ser que los criterios disminuyan. Sin conocer los criterios actuales es muy difícil mejorar la calidad y por lo tanto una empresa debe de tener una visión amplia sobre los criterios actuales.

Ajustando la calidad de datos
Por medio de un proceso deductivo y teniendo en mente el nivel deseado de calidad de datos seleccionados, esta tiene que ser mejorada paso a paso. Los ajustes pueden ser proactivos para los nuevos datos entrantes en el sistema o retrospectivos para los datos ya existentes. La mayoría de las empresas aplican las mejoras en la calidad de datos de manera proactiva, pero si se requiere hacer esta operación de manera retrospectiva, el éxito dependerá de la cantidad de esfuerzos involucrados comparado con su beneficio potencial. En algunas situaciones los datos no pueden ser mejorados, porque existen detalles que no son disponibles sobre ellos. En estos casos los beneficios no justifican los esfuerzos. No obstante, estos datos no podrán ser removidos de los recursos informáticos, porque siguen teniendo cierto valor. Para poder hacer frente a esta situación, tienen que ser establecidos diferentes niveles de certificación para los datos, como por ejemplo datos de alta calidad, datos de mediana calidad, y datos de baja calidad. La determinación de que datos se colocan en que nivel se toma por medio de un proceso de certificación de la calidad. Los datos que no cumplan con los criterios de un nivel pasan al nivel mas bajo hasta que encuentren su lugar. Cuando esto termine, los datos son certificados de acuerdo al nivel donde están. Este proceso es de mucha ayuda a la hora de identificar los datos de "mala" calidad dentro de los recursos oficiales de los datos.

Seguimiento a la calidad de datos

La trayectoria en los ajustes de la calidad de datos tiene que ser seguida para asegurarse que los criterios de calidad propuestos son aceptables y son aplicados de manera apropiada. En caso contrario se tienen que ajustar de nuevo. Este proceso cíclico de autocorrección continúa hasta que los criterios de calidad de los datos sean aceptables. Este proceso también puede ser usado de manera proactiva para identificar los datos que no cumplen con las características y cualidades necesarias, para estar en un nivel determinado. Se mejoran los procesos de colección y creación de estos datos para que se logre un resultado favorable y que de esta manera cumplan con todos los criterios. No existe ninguna medida absoluta para el nivel de calidad de los datos. Tampoco hay una escala de estándares que le determine o que califique la trayectoria de las mejorías en los datos. Esto es un problema, porque sin unas buenas medidas de calidad de datos y sin estándares que determinan dicha calidad, es difícil saber el nivel en que se encuentra la empresa y que si los pasos que se necesitan para crear los nuevos criterios son efectivos o no. Cuando se establecen los niveles de certificación los datos deben poder certificarse a cualquier hora. Si se mejora la calidad y los datos certifican cumpliendo con los nuevos criterios, automáticamente se pasa a un nivel mas alto de calidad. La cantidad de datos que pasa cada nivel de certificación son un indicador de los recursos de datos. Los cambios en la cantidad de datos que pasan a un nivel mas alto de certificación es un indicador viable que muestra la cantidad de datos mejorados. Aunque este enfoque parece subjetivo, en realidad sí proporciona una manera viable para medir tanto la calidad de los datos, como las mejoras y su efectividad.

CAPITULO 2
CALIDAD DE LA INFORMACION

2.1 Definiendo calidad de la información

Antes de poder mejorar la calidad de la información, debemos de definir su significado y como se puede medir. Vamos a definir que es y que no es la calidad de la información. Tenemos que definir también dato y información y sus conceptos claves, de la misma manera vamos a definir conocimiento y sabiduría, porque ahí es donde la información impacta al rendimiento del negocio y donde la información sin calidad lo perjudica. Durante la definición de la calidad de la información vamos a hacer diferencia entre calidad inherente y pragmática. Esencialmente calidad inherente es la correctitud de los hechos. La calidad pragmática es la correctitud de los hechos exactos representados correctamente. Vamos a terminar definiendo los tres componentes necesarios para la calidad de la información: calidad de la definición de los datos y su arquitectura, calidad del contenido de los datos, y calidad de la presentación de los datos.

2.1.1 ¿Qué es calidad?

La mejor manera de considerar la calidad de la información es ver que significa la calidad en el mercado en general y después hacer traducir que significa la calidad de la información. Como consumidores los seres humanos juzgan la calidad de las cosas de *manera consiente según su experiencia. La calidad es medida de manera consiente cuando* una persona compara los productos en el mercado y escoge uno de ellos como el producto buscado. "Buscado" significa que es el producto que mejor responde a sus necesidades y no necesariamente es el mejor en todas las categorías. Después de la compra la persona determina su calidad considerando si este producto a este precio responde a sus expectativas. La manera no consciente es la frustración que uno tiene de productos de mala calidad.

Primero vamos a definir que no es la calidad. La calidad no es un lujo o una superioridad y tampoco es lo mejor en su clase. La calidad existe en los ojos del consumidor basándose en su percepción o como responde a sus necesidades. Lo que para uno es calidad para otro puede ser algo defectuoso. Por ejemplo una compañía de seguros descargo sus datos para analizar los riesgos según los diagnósticos médicos de cada reclamo que fue pagado. Los datos revelaron que ochenta porciento de los reclamos tenían el diagnostico de "ruptura del pie". Que estaba pasando en realidad, los que procesaban el reclamo tenían como criterio pagarlo lo más rápido posible. De esta manera ellos dejaban en el sistema el diagnostico que aparecía por omisión que es "ruptura de pie". La calidad de la información es muy buena para pagar el seguro, porque el sistema requiere un diagnostico válido, pero para analizar los riesgos esta información fue completamente inútil. Que pasaría si un actuario determina el riesgo basándose en datos imprecisos. Que pasaría si las políticas del seguro determinaban el precio basándose en este riesgo. Que pasaría si el departamento de "servicio a clientes" enviaba una carta preguntando al cliente que tan bien se recupera de su "diagnostico médico". La información introducida para un solo propósito pero carente de calidad inherente va a truncar el crecimiento intelectual de la organización.

La calidad no es subjetiva o intangible. Ella se puede medir con la medición de negocio más fundamental - el impacto en los niveles más bajos de la organización

Entonces que es calidad. Podemos usar la definición de calidad como constante respuesta a las expectativas del cliente. Si una aplicación es diseñada y programada para responder al requerimiento funcional firmado por el usuario y durante las pruebas finales el usuario reclama que la aplicación no satisface sus necesidades esto significa que o las especificaciones del requerimiento o el análisis o el proceso de diseño fueron defectuosas. Calidad significa responder a las necesidades del cliente pero no necesariamente excederlas. La calidad significa que deberemos de mejorar las cosas que importan al cliente que hacen su vida más fácil y efectiva.

Tenemos que entender que son los datos, que es la información y porque se requiere saber sobre la calidad de la información. En el contexto de la ciencia de la computación el termino dato significa información numérica o alfanumérica representados en formato que la computadora puede procesar. No obstante podemos definir el dato desde el punto de vista del negocio independiente de la tecnología informática. Simplemente podemos decir que el dato representa los hechos sobre las cosas. Los datos representan cosas o entidades del mundo real. La entidad se puede definir como algo que tiene una existencia, separada o distinta y una realidad conceptual y objetiva. Cuando modelamos los datos estamos representando la clasificación de entidades que tienen características similares, como tipo de entidad El dato es un símbolo o otra representación sobre algún hecho o cosa. El tipo del hecho en realidad representa el tipo de atributo. El dato es el material crudo del cual deriva la información y es la base de acciones y decisiones inteligentes. Como ejemplo 00 3595456459 representa un hecho y es verdadero. Cuando esto representa algo real, pero falta la definición respectiva, este numero es sin significado. El dato es sólo el material crudo desde el cual se puede producir información.

2.1.2 ¿Que es información?

La información es un producto final. La información son datos en un contexto, que se pueden usar para un fin determinado. Es el significado de los datos, para que los hechos se vuelvan comprensibles. El ejemplo anterior se vuelve comprensible si sabemos que el número es un número telefónico, con la clave del país, la clave de la ciudad y el teléfono.
La calidad de la información requiere calidad de tres componentes: clara definición del significado de los datos, valores correctos y una presentación entendible (el formato de representación para el trabajador capacitado o el cliente). La baja calidad de cualquiera de estos tres componentes puede causar falla en el proceso del negocio o una toma de decisión equivocada. La información es un dato aplicado y se puede representar por la fórmula

Información = f(datos, definiciones, presentaciones)

Desde el punto de vista del negocio la información puede estar bien definida, los valores pueden ser precisos y pueden estar presentados de manera significativa, pero esto todavía no puede ser un recurso valioso para la empresa. La calidad de la información por si misma no tiene un uso, pero manejada y entendida por personas capacitadas puede llegar a tener un valor muy importante.

Conocimiento

La calidad de la información se vuelve un recurso poderoso que puede ser entendido y aplicado por las personas. Para que la información tenga valor se necesitan tanto trabajadores capacitados como calidad de la información. Una base de datos sin trabajadores con conocimientos que las están usando produce tanto valor como un data warehouse en producción sin clientes y pedidos. El conocimiento no es sólo conocer la información, esto es información en un contexto. El conocimiento significa entender el *significado de la información*. El conocimiento es información aplicada y se puede representar con la fórmula.

Conocimiento = F(gente , información , significancia)

El conocimiento es le valor agregado de la información por las personas que tienen la experiencia y la perspicacia de entender su potencial real. Con la continua evolución de la informática hoy el día las organizaciones son capaces de captar conocimientos en formato electrónico, organizarlo y compartirlo en toda la empresa. Los avances de Internet, Intranet y la minería de datos expanden el horizonte de datos compartidos en los data warehouses y en las bases de datos operativas.

Sabiduría

El objetivo de cada organización es de *maximizar* el valor de sus recursos informáticos para que cumplan su misión. El recurso informático es maximizado cuando es manejado de manera que tenga calidad y es accesible para los que lo necesitan. Los recursos humanos son maximizados cuando son capacitados y tienen recursos incluyendo informáticos y son empoderados para actuar y cumplir con el trabajo de la empresa y satisfacer al usuario final. La sabiduría es conocimiento aplicado y se puede expresar con la formula.

Sabiduría = f(personas , conocimiento , acción)

El objetivo de la calidad de la información es proveer a los trabajadores capacitados con los recursos estratégicos para llegar a formar una organización inteligente. Una organización inteligente continuamente expande su capacidad de crear su futuro mediante el acceso, la lectura y el aprendizaje compartido. La organización inteligente maximiza su experiencia y sus recursos informáticos en el proceso de aprendizaje. Ella comparte *abiertamente* la información entre sus distintas áreas y de esta manera crecen conjuntamente tanto la organización como los trabajadores.

2.1.3 ¿Que es calidad de la información?

Existen dos definiciones significativas de la calidad de la información. Una es su calidad inherente y la otra es su calidad pragmática. La inherente es la validez y la pragmática es la correctitud de los datos. La pragmática es el valor que estos datos, correctos y validos, tienen para el trabajo de la empresa. Datos, que no ayudan a la empresa para cumplir su misión no tienen calidad sin importar que son correctos o validos.

Calidad inherente de la información.

La calidad inherente es simplemente precisión. Es el grado de precisión con el que los datos reflejan el mundo real que representan. Todos los datos son una abstracción o una representación de algo real.

Calidad pragmática de la información.

La calidad pragmática representa el grado en que los datos son útiles y valiosos para apoyar el proceso de la empresa, para cumplir con sus objetivos. En esencia la calidad pragmática es el grado de la satisfacción del cliente, derivado por los trabajadores capacitados, que usan la información para cumplir con su trabajo. Los datos en una base de datos o en un data warehouse no tienen valor actual sino potencial. Los datos tienen su valor realizado cuando alguien los usa para hacer algo útil. La calidad pragmática es el grado en que los datos ayudan al trabajador para cumplir con el objetivo de la empresa de forma eficaz. La calidad pragmática previene los trabajadores de: hacer mal su trabajo o tomar decisiones equivocadas, hacer dos veces el mismo trabajo, corregir el impacto de una mala decisión o perder tiempo innecesario para investigar la integridad de los datos antes de usarlos. Efectuar cálculos o reformateo de los datos antes de usarlos, buscar información adicional con el objetivo de entender y de usar los datos implica perder tiempo y clientes, causar daños irreparables, perder oportunidades de negocio, falta de comunicación interna y con los clientes.

2.1.4 Componentes de la calidad de la información.

En teoría la información esta ligada y en función de tres componentes principales. Estos componentes son los datos y su arquitectura, su definición y su presentación. Ellos determinan que tan completo es cualquier producto informático. También son separados en distintos subcomponentes, cada uno de los cuales tiene su propia calidad de la información. Si no se conocen la definición o los hechos representados por los datos, entonces cualquier valor carecerá de un significado útil y por lo tanto no tendrá calidad de la información. Lo mismo pasara si se conoce la definición, pero el valor es incorrecto, en este caso tampoco se obtiene calidad. Por otro lado si todo estuviera en orden, pero se carece de presentación, entonces el técnico capacitado no sabrá como interpretar la información y tampoco tendríamos calidad.

2.1.4.1 Calidad de los datos y su arquitectura

La definición se refiere a las especificaciones de un dato que constan de un conjunto de valores del dominio y las reglas que rigen el dato. La calidad de definición de un dato es el grado al que se define este dato· describe de manera exacta el significado de algún hecho del mundo real, que el dato representa y cumple con las necesidades de todos los usuarios de la información, para que puedan entender el dato que están usando. Dentro de los clientes de la información se incluyen los clientes del negocio y el personal de sistemas de información. Los siguientes grupos de personal son siempre usuarios de la información y la manera como están involucrados es·

- Los trabajadores capacitados deben de conocer el significado de la información para poder cumplir con su trabajo
- Los capturistas y creadores de la información deben de entender su significado junto con los valores válidos y las reglas del negocio que los rige para poder crearlos o mantenerlos actualizados.
- El personal que administra los datos debe de conocer el significado de la información con todos sus atributos, para poder desarrollar modelos precisos de los datos, y también para poder diseñar bases de datos de alta integridad.
- Los analistas de sistemas deben de conocer los puntos mencionados anteriormente, para poder desarrollar aplicaciones de modelos de alta integridad
- Los desolladores de aplicaciones necesitan estos puntos para poder proporcionar aplicaciones lógicas de alta integridad.

La calidad de la arquitectura de la información es el grado en el cual la estructura de los datos:

- Implementa las relaciones reales y inherentes de los datos, para representar de una mejor manera los eventos y los objetos del mundo real.
- Es estable, permitiendo la incorporación de nuevas aplicaciones, reutilizando los mismos datos originales sin modificación. De esta manera los nuevos datos introducidos no tendrán redundancias con respecto a los viejos datos y como resultado únicamente se dará el caso de introducción de nuevos atributos dentro del modelo de dato o las bases de datos existentes. Estabilidad de la base de datos significa que las nuevas aplicaciones pueden usar los datos actuales y que estos no tendrán que sufrir alguna modificación o cambiar de estructura dentro de la base de datos, simplemente añadiendo los datos nuevos.
- Es flexible, soporta nuevos cambios dentro de los procesos de la empresa sin cambios significantes al modelo de datos o las bases de datos. Una base de datos flexible significa que dos líneas de negocios se pueden integrar para que se elimine la duplicidad y para que se maximizen las ventas con cambios mínimos en el diseño.

Una definición clara y precisa es requerida para asegurar una comunicación clara entre todos los que manejan la información. Igual como un diccionario lexicográfico define el significado de las palabras, de la misma manera la empresa requiere una lexicografía para el negocio donde se definen de manera exacta los términos del negocio y los hechos. Pueden existir diferentes términos de negocio en diferentes contextos, por esto cada definición debe de ser incluida en un glosario empresarial. La calidad de la definición de los datos es una característica y una medida de los modelos de datos implementados y desarrollados por el área de sistemas.

2.1.4.2 . Calidad del contenido de los datos

La calidad de la información requiere calidad de la definición de datos y de contenido de datos. La calidad del contenido de los datos es el grado en el cual los valores representan con precisión las características del mundo real o de los procesos y responden a las necesidades reales de información por parte de los usuarios. Le empresa debe tener una representación clara sobre el cliente llamado Pérez García Jesús Mauricio, para que tenga

una relación eficaz con él y que conozca todos los productos contratados por él, para definir la tendencia de sus necesidades y la venta de nuevos productos a este cliente. La calidad del contenido de los datos es una característica y medida de los datos creados y actualizados durante los procesos de negocio y de las aplicaciones que los implementan.

2.1.4.3. Calidad de la representación de los datos.

Aun si los datos y la arquitectura estén bien definidos y tienen contenido preciso el proceso de negocio todavía puede fallar. Las causas posibles de las fallas pueden ser:

- Cuando el dato es inaccesible
- Cuando el dato no es disponible a tiempo y de manera oportuna
- Cuando el dato es representado de manera ambigua con una etiqueta la cual es inconsistente con la definición y el nombre del dato.
- Cuando la presentación del dato requiere un trabajo adicional para su interpretación
- Cuando el dato viene combinado con otros datos de manera incorrecta, produciendo así datos derivados incorrectos.

La calidad de representación se aplica a documentos y medios de entrega como son los reportes o las ventanas del resultado de una búsqueda en las bases de datos o en el data warehouse. La representación de los datos tiene que ser enfocada hacia las necesidades de los trabajadores capacitados y que ellos puedan entender fácilmente el significado y la explotación de la información. Debido a que la información se usa para diferentes propósitos se necesitan diferentes formatos de la presentación. En este caso su calidad requiere que el formato de presentación sea intuitivo. La calidad de la presentación de los datos es una característica y medida de cómo la información es accesada por los trabajadores calificados y cual es su presentación.

Representado de otra manera, la calidad da la información y el beneficio que esta brinda a los trabajadores capacitados, aparece en la tabla 2.1.

Tabla 2.1 La calidad de la información y el trabajador capacitado.

Característica de la calidad	**Beneficio para el personal**
El dato correcto	El dato que yo necesito
Con la completes correcta	Todos los datos que yo necesito
En el contexto correcto	Cuyo significado yo conozco
Con la precisión correcta	Yo puedo confiar en el
En el formato correcto	Lo puedo usar fácilmente
En el momento correcto	Cuando yo lo necesito
En el lugar correcto	Donde yo lo necesito
Para el propósito correcto	Yo puedo cumplir con mis objetivos y satisfacer al cliente.

2.1.5 Aplicación de los principios de calidad a la información.

Aplicar los principios de calidad a la información significa enfocarse en los clientes de la información, en los productos informáticos creados, y los procesos que crean, actualizan y presentan la información.

La calidad tiene que ser enfocada a los usuarios de la información

No se puede mejorar la calidad de la información sin antes pensar en quienes son nuestros verdaderos usuarios, y si lo que necesitan está en los productos informáticos.
Los usuarios informáticos constan de:

Usuarios finales. Las empresas no sólo proveen de productos y servicios a sus usuarios, sino que también de productos informáticos por medio de generación e envío de información. Esto puede ser por medio de catálogos, anuncios, sitios WEB, y correspondencias.

Personal interno. Todos los empleados que requieren de la información para hacer su trabajo son usuarios de la información.

Personas externas involucradas con la actividad de la empresa. Esto incluye accionistas y consejeros que necesitan información sobre sus inversiones, autoridades que necesitan información para asegurar que la empresa cumple con las leyes.

2.1.5.1 Los datos son la materia prima – la información es el producto final.

Los datos son la materia prima producida y modificada por uno o más procesos del negocio. Estos datos llegan a ser el material para otros procesos del negocio.
El flujo de la información es el conjunto completo de procesos y aplicaciones computarizadas los cuales crean, actualizan, extraen, interrelacionan, transforman datos y presentan la información desde su inicio original o su creación de conocimiento en formato electrónico o algún otro formato. Esta información presentada esta utilizada para búsquedas y una visualización por los usuarios finales. El flujo de información también incluye los suministradores de la informaron, accionistas, consejeros y organizaciones externas involucradas. El flujo de la información finalmente incluye todas las bases de datos y archivos (sin importar que son manuales o computarizados) como son librerías, bibliotecas de deportes donde los datos son almacenados desde su origen hasta el sitio oficial donde se almacenan. *El flujo de información expresa la relación informacional entre usuario* y suministrador. La cadena de suministro de información esta presentada en la figura 2.1.

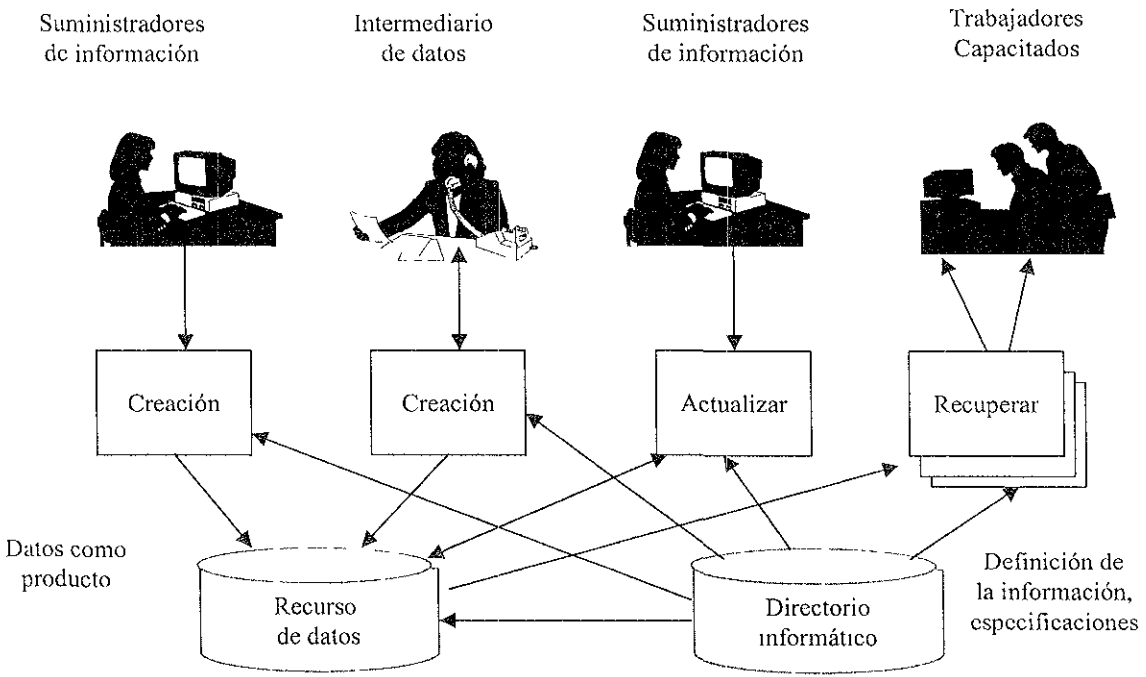

Figura 2.1 El flujo del producto informático

El proceso implica la existencia de roles de "suministrador" y "usuario" de la información. La persona que origina o actualiza los datos es el "suministrador" o "productor de información". El "usuario" es cada uno del personal que usa la información en su trabajo.
Así como en un producto de manufactura se aplican los principios de la calidad, los mismos principios se pueden aplicar al proceso de generación y actualización de la información. En este caso es importante hacer diferencias entre el suministrador de la información y el intermediario de datos. Un intermediario es alguien que toma los datos de una forma y los transcribe en otra, aportando un valor agregado mínimo. Por ejemplo, un empleado que atiende llamadas telefónicas recibe la llamada de un cliente y llena una forma con la información obtenida. Después esta forma pasa a los capturistas de datos y ellos introducen la información desde la forma a la base de datos Los capturistas son intermediarios de datos. El intermediario de datos es un punto adicional en el manejo de los datos y puede provocar demoras en el acceso a los datos, o empeorar la calidad de la información cometiendo errores. El error fundamental en la definición de los procesos hoy es que se están definiendo únicamente los productos tangibles o físicos, mientras la información esta tratada como subproducto, o documentación o simplemente papeleo. En la época de la informática ninguna empresa puede explotar el valor de sus recursos informáticos sin una clara visión que la información es un producto fundamental del proceso del negocio. Los datos reunidos sobre la calidad de los procesos y eventos del negocio se pueden analizar, para descubrir tendencias y perspicacias imposibles de detectar de otra manera.

2.1.5.2 Planeando y obteniendo calidad de la información.

La calidad de la información requiere de una cuidadosa planificación y una controlada ejecución. La planificación de la calidad de la información incluye la definición de cual información es requerida para que el negocio funcione correctamente y cumpla con su misión. Las actividades se aplican a los dos, el negocio mismo y el desarrollo de sistemas de la empresa, bases de datos y data warehouse.

Los procesos de producción de la información son simplemente los procesos de negocio, incluso de manufactura en los cuales se esta creando, coleccionando, capturando sin importar si esto se hace de manera manual o electrónicamente. Cada proceso incluye personal responsable para el trabajo y su producto final incluyendo el producto informático.
Los roles en el proceso de calidad de la información, descritos de manera breve, incluyen:

- Trabajadores capacitados a nivel estratégico y táctico
- Trabajadores operativos
- Suministradores de la información
- Intermediarios de los datos
- Gerentes, arquitectos, analistas de la información.
- Diseñadores y administradores de la base de datos
- Analistas del negocio y de los sistemas
- Diseñadores de las aplicaciones y desarrollos
- Diseñadores del datawarehouse y arquitectos

(Ver figura 2.2)

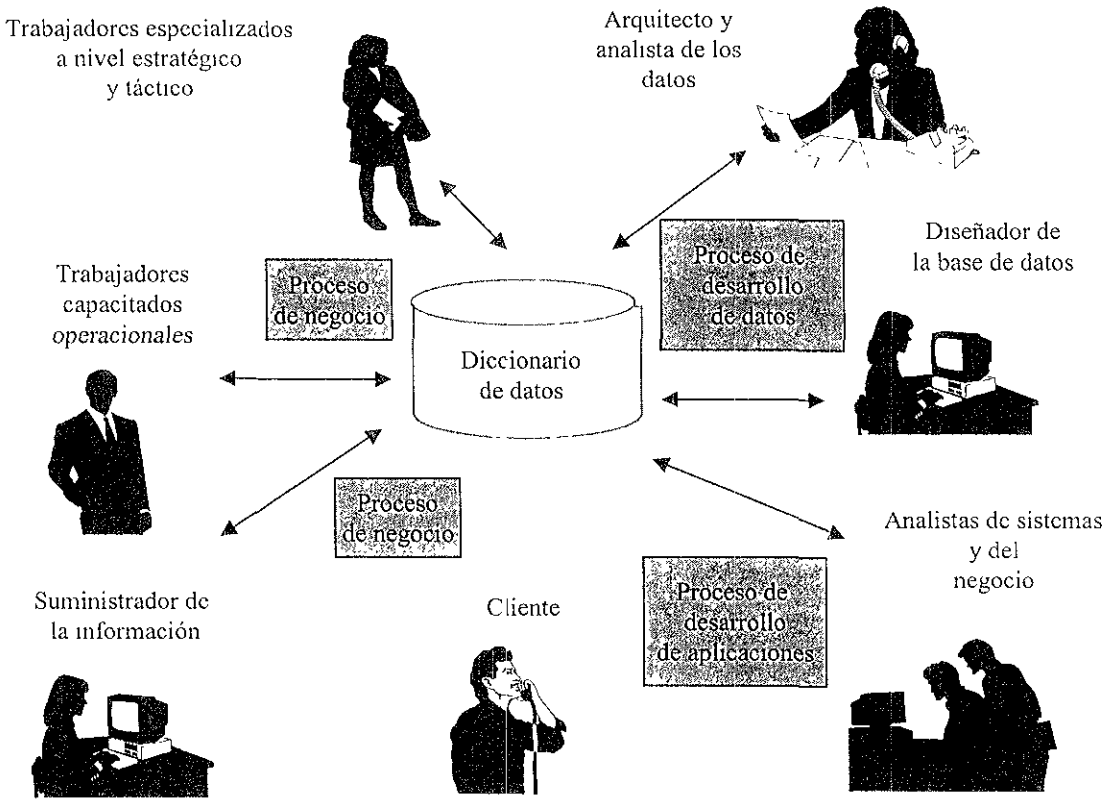

Figura 2.2 Componentes en la planeación de la calidad de información

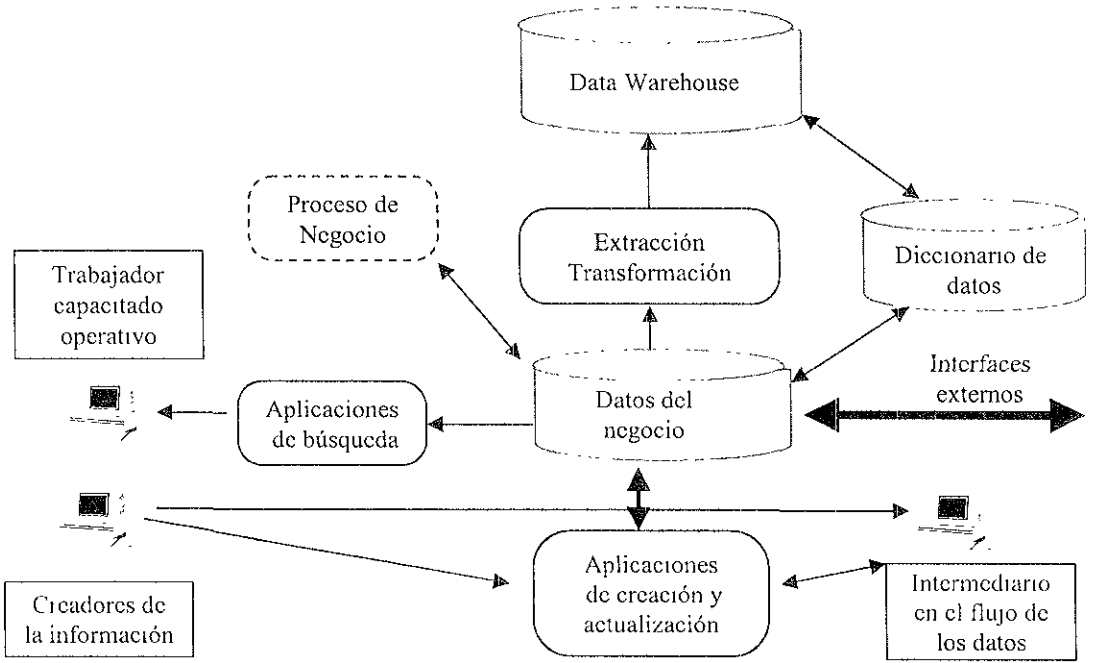

Figura 2.3 Componentes de la calidad de la información dentro del proceso de negocio.

Los trabajadores capacitados a nivel táctico y estratégico

Este es el punto de inicio para la definición de los datos y la calidad de la información. Si una empresa no empieza por este punto, para planear y desarrollar su data warehouse, su proyecto fracasará. El data warehouse es diseñado para apoyar a los trabajadores capacitados de nivel táctico y estratégico, que soportan los procesos de la empresa a nivel táctico y estratégico respectivamente. La información en el data warehouse tiene que ser definida de tal manera, que pueda soportar las bases, los requerimientos, así como las necesidades de información derivada y sumarizada. Los trabajadores de nivel táctico y estratégico y los administradores de la información tienen que ser incluidos también en el proceso de definición de los datos y la calidad de la información en la labor de planificación y desarrollo de las aplicaciones operacionales y las bases de datos. Si esto no sucede de esta manera las aplicaciones y bases de datos resultantes no podrán adquirir un valor máximo para la empresa. La razón de esto es que si no se consideran los requerimientos estratégicos de la información, esto puede provocar que los procesos de análisis y diseño no puedan definir aspectos claves, tanto objetos como eventos, para el negocio. Estos son claves para poder realizar a cabo el análisis de las tendencias de los procesos, pero no afectaran los procesos operacionales. Cuando se analizan los requerimientos operacionales para cualquier aplicación operacional, siempre se incluye una visión de la información (búsquedas, decisiones, e indicadores claves del negocio) que se requieren, para apoyar los procesos estratégicos y tácticos. Esta información aveces no es necesaria para soporte de los procesos operacionales inmediatos, pero se necesita para la toma de decisiones. Esto minimiza la posibilidad de pasar por alto atributos claves, que tienen que ser capturados por procesos automatizados.

Trabajadores operativos

Como los trabajadores operativos realizan los procesos de negocio de la empresa, cada aplicación y base de datos, tiene que ser diseñada, para capturar la información y de esta manera cumplir con sus necesidades. Esto incluye todos los trabajadores capacitados de los departamentos a través de los cuales pasa la información. Para un diseño con calidad de la información, los datos tienen que ser definidos de manera consistente de tal modo que cualquier trabajador que los use los tenga completos. Obviamente esto va a incrementar el trabajo del empleado que suministra los datos, pero disminuirá la carga de todos los demás trabajadores, que se encuentran en la cadena. Cuando se definen los requerimientos de información para una aplicación nueva, el análisis debe de identificar todos los departamentos y los trabajadores capacitados, que tengan algo que ver con el flujo de los datos. Estos últimos se deben de involucrar activamente en el proceso de definición de los datos. Considerar todo el proceso y la información necesaria, para los trabajadores, para que el modelo de información diseñado sea completo.

Suministradores de la información

Los suministradores de la información son claves en la cadena de suministro con información Como fuente principal de información, también son de suma importancia para su calidad. Los suministradores se tienen que involucrar en la definición de la información

que se va a automatizar, junto con los trabajadores operativos de tal manera que la definición debe ser acordada entre ambos, el suministrador y el usuario de la información.
Los suministradores tienen que saber quienes son sus usuarios y cuales son sus necesidades Los trabajadores operativos tienen que conocer los que les proporcionan la información y los problemas eventuales, que pueden aparecer en el proceso de captura. De esta manera se puede consolidar un trabajo en equipo, en lugar de una competencia entre ellos, que permita un mejor desempeño de la empresa.

Intermediarios de los datos

Los intermediarios son los que transcriben los datos de una forma a otra. El oficinista que introduce los datos desde el papel a la base da datos tiene un rol de intermediario. A pesar que los intermediarios no definen los requerimientos de transformación de datos, ellos también son trabajadores capacitados. Ellos usan otra información en el proceso de introducción de los datos y por esto deben de tener participación en la definición de los datos, ya que tienen que saber que necesitan para hacer su trabajo. Igual deben de participar en el proceso de diseño de la aplicación, para minimizar ciertos tipos de problemas de calidad.
Hay que tratar de eliminar intermediarios de datos que no son necesarios. La calidad es mejorada cuando la información se captura de manera electrónica o por los suministradores. Los intermediarios no pueden corregir los errores encontrados, porque no crean la información, y por esto los errores tienen que ser guardados en un archivo de excepciones, para su futura limpieza.

Gerentes, arquitectos y analistas de la información

Ellos son responsables para la integridad de la estructura de la información y el modelo de datos. Ellos construyen la definición junto con los expertos en asuntos de negocio con una clara y robusta definición, relaciones precisas, dominio de valores y reglas de negocios completos. Su responsabilidad es de crear una estructura estable, flexible y reutilizable. El propósito del modelo de datos es de mejorar el negocio, la productividad, y las comunicaciones entre los sistemas.

Diseñadores y administradores de la base de datos

Ellos transforman el modelo lógico de datos en un modelo físico. Su responsabilidad es la integridad de diseño físico, balancear el rendimiento, asegurar la integridad física, mantener la seguridad de acceso y recuperación en caso de fallas.

Analistas de negocio y de sistemas

Los analistas son los que analizan y definen requerimientos, para las aplicaciones. Personas que son especialistas en las funciones del negocio, ocupan estos puestos y efectúan definiciones que involucran varios procesos, dentro del negocio. Ningún proceso puede ser definido sin especificar su producto final. Ninguna especificación de una definición es completa sin contener el producto final de la información. Su responsabilidad es facilitar la

definición del proceso de negocio y crear un requerimiento transparente que pueda ser convertido en una aplicación automatizada. El analista debe de preguntarse quienes son todos los usuarios del producto final del proceso. Estos pueden ser trabajadores capacitados internos o clientes externos. Cuales son las expectativas reales, para el producto en este proceso. Que debe de hacer el proceso para cumplir con estas expectativas. Los analistas deben de ser capaces de leer entre las líneas, para distinguir entre el requerimiento "expresado" por los involucrados en el proceso y al requerimiento real de los que usan el resultado de este proceso.

Diseñadores de las aplicaciones y desarrollos

Son las personas que diseñan, desarrollan y implementan las aplicaciones automatizadas para todo o parte del proceso del negocio. El diseño de la presentación de la información es de suma importancia. Ellos diseñan pantallas, formatos de reportes, gráficas, para visualizar datos y otras formas de presentación de la información. Su responsabilidad es asegurar el diseño y la implementación de las aplicaciones, con programas que cumplen con las especificaciones del requerimiento. En el desarrollo se deben de involucrar los trabajadores que realizan el proceso junto con los trabajadores que son usuarios de la información producida por este proceso. Ellos deben de ser incluidos en forma proactiva en los prototipos del diseño y el desarrollo. De esta manera la fase de pruebas empieza durante el desarrollo.

Diseñadores del data warehouse y arquitectos

Ellos son los encargados del diseño y la planeación del data warehouse, los datamarts, o los sistemas de información, para los ejecutivos. La diferencia principal entre los arquitectos del data warehouse y los sistemas de producción radica en la naturaleza de los procesos que soportan. Las bases de datos del data warehouse tienen que ser modeladas y diseñadas para soportar los procesos tácticos y estratégicos. Las bases de datos de producción tienen que ser diseñadas primero para soporte de los procesos operativos y después de los estratégicos y tácticos. La base de datos con los datos operativos almacenados tiene que ser diseñada con un consenso de toda la empresa sobre su definición. En un ambiente de datos fragmentado esta se puede convertir en una base de datos operativa, que contiene referencia *a las múltiples ocurrencias de entidades de datos, proveniente de varios aplicativos* heredados. Existen ciertos eventos de negocio que se pueden conocer únicamente durante el transcurso de una transacción de negocio. Lo principal para un data warehouse es que debe de ser parte de la definición común de la arquitectura de la información de la empresa.

2.1.5.3 La calidad de la información – un servicio al cliente.

Si interpretamos la información como un producto esto llevaría al concepto de calidad de la información a un concepto de servicio al cliente. Debido a que los trabajadores capacitados necesitan información con calidad para cumplir con sus tareas, ellos dependen de los suministradores de la información. Si consideramos que la información es un producto y los trabajadores son los usuarios, entonces proveer información con calidad significa dar servicio a estos usuarios. La creación de la información es un proceso de "manufactura", en el dominio de los datos necesarios para efectuar el trabajo "físico" y de "conocimiento" de

la empresa En una empresa con calidad los dueños de los procesos y los suministradores de la información siempre deben de tomar en consideración todos los trabajadores involucrados en la cadena de la información.

2.2 Herramientas para la calidad de la información.

Las herramientas para la calidad de información son efectivas si no se abusa de ellas. En un ambiente apropiado, donde existe un respeto hacia la satisfacción de las necesidades del usuario, combinados con un conjunto de procesos de calidad de la información, las herramientas pueden ser usadas de manera muy efectiva, para incrementar la eficiencia y el valor de la calidad de la información, de la limpieza de datos, controlarlos y mejorarlos. Si no existe una cultura de la calidad de información, las herramientas no serán útiles, o a lo mucho servirán, para absolver la propia responsabilidad de los encargados de la calidad y la limpieza de datos. Existen varias categorías de productos relacionados con la calidad de información con sus respectivos enfoques, problemática y herramientas. Sólo nos enfocaremos a las principales funciones que las herramientas proporcionan, que ayudan para sobresaltar los criterios para la evaluación de los productos de limpieza de datos. También son importantes las técnicas, para mejorar la calidad de la información y los pasos, que los procesos deben seguir, para alcanzarla.

Las herramientas para la calidad de la información automatizan los procesos y proporcionan soporte, para ofrecer soluciones a los problemas de calidad de la información. Para lograr un uso efectivo de las herramientas de limpieza de datos se requiere de los siguientes puntos básicos:

- Entender el problema que se esta resolviendo
- Entender el tipo de tecnologías disponibles y su funcionamiento en general.
- Entender el alcance de las herramientas.
- Entender sus limitaciones
- Saber escoger la herramienta correcta en base a las necesidades
- Usar las herramientas de manera apropiada

Existen cinco categorías de productos informáticos que se pueden usar para mejorar la calidad de información. Estas categorías se acomodan dentro de la cadena de flujo de información representados en la figura 2.4.

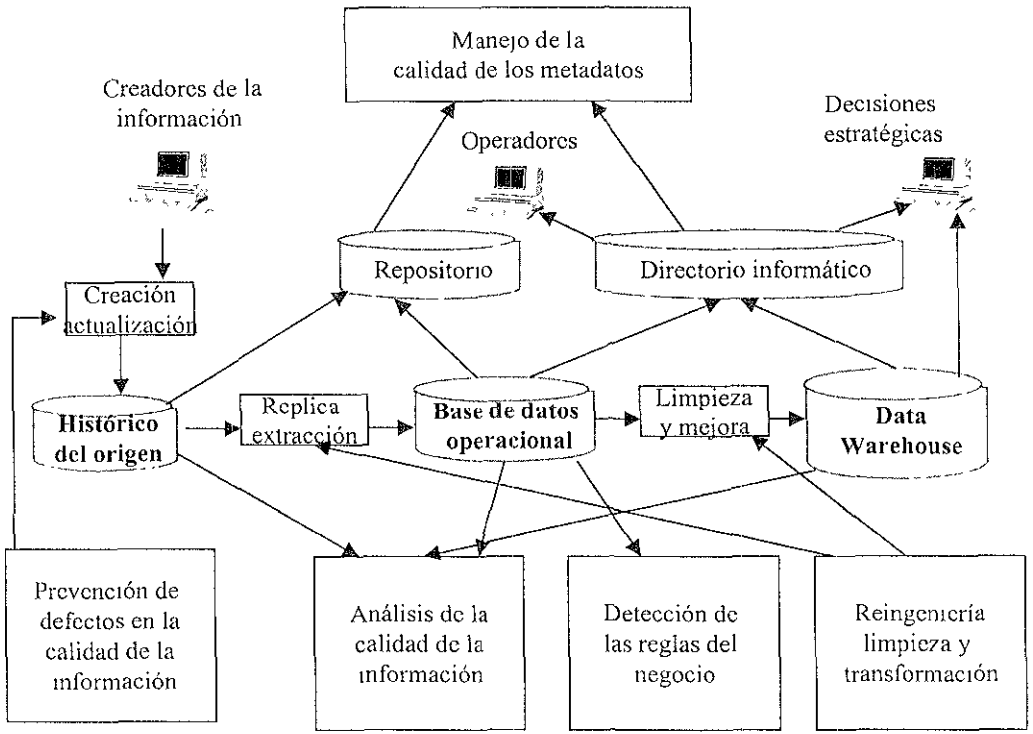

Figura 2.4 Las herramientas de la calidad de la información relacionadas con el flujo de información

Categorías de las herramientas para la calidad de la información.

La clasificación de las herramientas que proporcionan una mejoría de la calidad de información se muestra a continuación:

- Herramientas para el análisis de la calidad de información. Las herramientas de análisis extraen los datos desde una base de datos o algún proceso, miden su calidad, comparando su validez según las reglas de negocio establecidas, y generan reportes de su análisis.
- Herramientas para detección de las reglas de negocio. Este tipo de herramientas analiza datos, para descubrir los patrones de comportamiento y las relaciones que existen entre ellos mismos. De esta manera se logran identificar las reglas del negocio, por medio de un análisis de los patrones de conducta de los datos.
- Herramienta para la reingeniería, limpieza y transformación de datos. Las herramientas, para la corrección de los datos extraen, estandarizan, transforman, corrigen, (donde sea posible) y mejoran los datos, sea en su lugar de origen o en su camino, hacia el data warehouse.
- Herramientas para prevención de fallas en la calidad de datos. Estas herramientas evitan que los errores entren en la base de datos o que existan violaciones de las reglas del negocio. Programas aplicativos, que crean y actualizan los datos llaman a módulos o rutinas para la prevención de defectos. Estas herramientas aplican las reglas de negocio y efectúan pruebas de la calidad de los datos durante el proceso de su creación y actualización.
- Herramientas para la calidad de los metadatos. Estas herramientas controlan las definiciones de las reglas de negocio, las reglas de transformación de los datos o evalúan y controlan el mismo metadato, como por ejemplo si cumple con el estandarte de los nombres.

La mayoría de los productos para calidad de los datos proporcionan las funciones de mas de una categoría. Por ejemplo, un producto puede encontrar las reglas del negocio y después reportar las variaciones fuera de las reglas. Algunas herramientas de limpieza de datos que limpian y transforman datos en las bases da datos existentes, también tienen módulos o rutinas que pueden ser llamadas y ejecutadas directamente desde programas que crean y actualizan datos.

En la tabla 2.2 se listan las diferentes clasificaciones de herramientas para la calidad de información. El código de clasificación de la tabla será usado en las tablas posteriores, para la calificación de las diferentes herramientas. Este código identifica las cinco categorías mencionadas anteriormente. También se identifica cuando una herramienta de reingeneria y limpieza es una herramienta general (CG), la cual puede ser usada, para la transformación o corrección de cualquier tipo de dato. Otras son dedicadas específicamente para los nombres y direcciones de los datos (CN). Algunos de los proveedores cuentan con servicios de terceros, además de vender los productos (S), otros ofrecen únicamente servicios.

Tabla 2.2. Clasificación de las herramientas, para la calidad de la información.

Código de clasificación	Nombre de la clasificación	Descripción de la clasificación
A	*Análisis*	*Evaluación automatizada de los datos*
C	Limpieza	Extracción, reingeniería, transformación, y/o limpieza de los datos.
CG	Limpieza de tipos generales de datos	Reingeniería, limpieza, y transformación de datos de cualquier tipo.
CN	Limpieza de nombres y direcciones	Limpia y mejora datos de nombre y dirección
M	Calidad de Metadatos	Proporciona validaciones sobre los metadatos.
P	Prevención de datos *defectuosos*	Evita errores en aplicaciones donde se originan y *actualizan los datos.*
R	Descubre las reglas	Analiza y descubre las reglas del negocio en los datos.
S	Proporciona servicios	El proveedor proporciona servicios de calidad o tiene otros proveedores que usan el producto para ofrecer el servicio de calidad.
SO	Solo proveedor de servicios	El proveedor solo proporciona el servicio de calidad de información y no vende el producto.

A *continuación vamos a describir las funciones básicas de cada categoría de herramientas* y presentaremos lista de productos de cada categoría.

2.2.1 Herramientas para el análisis de la calidad de la información.

Las herramientas automatizan parte de la valoración de la calidad de información y los procesos de auditoria. También automatizan la medición de la calidad de los datos.

Funcionalidad de las herramientas.

Las funciones de valoración prueban si los datos se comportan según su definición y la de la regla del negocio. El software para análisis de calidad proporción una valoración automatizada. Este software generalmente solo requiere que se definan las reglas del negocio en su repositorio. El software convierte las reglas introducidas en código para validar las reglas definidas contra los datos. Algunas herramientas pueden encontrar solas las reglas del negocio a partir de los datos mismos e identificar aquellos que no cumplen con la regla

Los productos de análisis deben medir las siguientes características de la calidad de la información:

- Completes de los valores
- Validaciones conforme a las reglas del negocio como:
  - Validez del dominio de valores
  - Valida si el dato pertenece a un rango de valores
  - Integridad referenciál

  - Autenticidad de llaves primarias únicas
  - Operaciones con datos derivados
- Precisión de los valores de los datos
- Equivalencia de los datos redundantes en múltiples bases de datos

Las funciones de las herramientas de análisis de calidad son:
- Extracción de datos, incluyendo extracción al azar
- Definición de las reglas del negocio y especificaciones para la medición de la calidad
- Automática valoración de los datos comparada con las reglas del negocio
- Emitir reportes con los resultados descubiertos sobre los datos de diferentes formas.
  - Reportes de excepciones
  - Por medio de mapas, tablas, columnas o porcentajes y conteo relativo de los errores
  - Gráficas y el análisis del costo de los errores
  - Mapas de control que muestran la historia de los valores en el tiempo

Las cualidades y los puntos fuertes de las herramientas de software para la calidad de datos son que pueden desarrollar funciones automáticas de manera rápida y precisa. La capacidad de integrar varios reportes ahorra tiempo en la presentación del análisis. El repositorio se convierte en una fuente para documentar las reglas de negocio y puede ser compartido entre todos los trabajadores involucrados. También aveces existen algunas limitaciones, pero en este caso el administrador encargado de la calidad debe de conocerlos previamente.

Limitaciones de las herramientas.

Los productos de análisis miden únicamente la validez y la completes, ellos no pueden medir la mayoría de los tipos de precisión Por ejemplo, una dirección puede ser válida, pero la persona puede haber cambiado de vivienda. El precio de un producto puede ser dentro de los límites razonables, pero el precio podría estar no actualizado. Estos son ejemplos de validez, pero no de precisión. Una herramienta de análisis puede identificar que la distribución de los valores es sospechosa, pero no puede definir si es correcta o no. Entonces cuando las herramientas leen los reportes de los análisis deben de hacer diferencia entre precisión y validez Los trabajadores capacitados no deben de asumir que los datos que pasaron la prueba de validez son precisos.

En la tabla 2.3 se presentan los productos que efectúan análisis de la calidad

2.2.2 Herramientas para la detección de las reglas de negocio.

Las herramientas para la detección de las reglas del negocio sirven para entender como se usan los datos. Estos son unas herramientas especializadas en minería de datos y se enfocan en el entendimiento de las reglas de negocios vigentes. Las herramientas soportan preparación de los datos para limpieza

Tabla 2.3 Proveedores y productos representativos que ofrecen software, para el análisis de la calidad de la información.

Proveedor	Nombre del producto	Código de clasificación
Data Flux Corp.	Data quality Workbench, SmartScrub, Datalogue, Extend, MatchMaker, IntelliMerge, Inspector	CG, A, R, CN
DBE Software	DB- Examiner	AM, PM
Decisionism, Inc	Aclue Decision Supportware	A
DupeKiller, Inc	Dupe Killer	A, CN, S
Gamma Research	OCRProof	A, P
Gladstone Computer Sevices	DQ Administrator, Warehouse Quality Administrator	CG, A
Innoative Systems	Analyzer, Verify, Dictionary, Edit, Match, Scrub, Household, CorpMatch, Find	CN, CG, A, M, R, P, S
MatchWare Tecnilogies (a subsuduary of Valıty)	AutoStan, AutoMatch, MatchWare/CL, MatchWare/ PACE	A, R, C
Mobius Inc.	INFOPAC-ABS	A, R
OTS Group	Global Thirth Party Name/adress, Cleansing & Enhancment Service	A, CG, CN, SO
Pine Cone Systems	Content Tracker, Refreshment Tracker	A, M
Prism Solutions (adquired by Ardent Software, Spring 99)	Prism Quality Manager	C, G, A
Rockwell Automation DataMyte	*Quantum SPC/DC, Quantum* SPC/QA	A
Search Software America	NAME 3, EXTENSIONS, Data Clustering Engine	A, R, CN, P
Unitech Systems	ACR/Plus(Detail Summary Data)	R, A
WizSoft	WizRule, WizWhy	R, A

Funcionalidad de las herramientas.

Las herramientas analizan los datos por campos, archivos o múltiples archivos, para detectar patrones de relaciones y reglas. Ellas usan la minería de datos o algoritmos de inteligencia artificial, para analizar los datos y deducir las reglas de negocio aplicadas, para el manejo de la calidad de la información. Las herramientas analizan varios tipos de patrones y reglas como sigue:

- Contadores de valores en un dominio
- Frecuencia de la distribución de los valores de datos
- Patrones de valores en datos compuestos (como son los datos de textos, direcciones o nombre)
- Formulas o algoritmos de calculo
- Relaciones como son los datos duplicados entre los archivos
- Similitud de elementos
- Correlación entre valores de datos en diferentes campos como son la colonia y código postal
- Patrones de conducto que puedan indicar posibles fraudes, tanto intencionales o no intencionales

Limitación de las herramientas.

Generalmente las herramientas encargadas de descubrir las reglas del negocio no siempre son capases de descubrir todas las reglas. Las reglas pueden ser mas de lo que es capaz de advertir el algoritmo que las examina. También puede darse el caso de que algunas reglas descubiertas pueden ser impertinentes. Optimizar los parámetros de búsqueda de la herramienta puede minimizar esta situación. Por otro lado pueden existir problemas de rendimiento con algunos archivos muy grandes o cuando se analizan demasiados campos. Este caso puede ser corregido por medio de muestras aleatorias y haciendo análisis separados de los diferentes conjuntos de campos, agrupadas de una manera lógica, para que las reglas del negocio puedan ser distinguidas de manera fácil.
Siempre es conveniente asegurarse que todas las variables sujetas a las reglas del negocio son disponibles, para el análisis. Es bueno excluir del análisis, campos que pueden ser irrelevantes, para el tipo de regla que se busca encontrar.

En la tabla 2.4 se presentan los productos que efectúan la detección de las reglas del negocio.

2.2.3 Herramientas para reingeniería y limpieza de los datos.

Las herramientas para reingeniería de datos y las herramientas para limpieza se usan para mejorar la calidad de los datos en si. Proporcionan automatización de los procesos de corrección de datos.

Funcionalidad de las herramientas.

Las herramientas para reingeniería y para limpieza de datos pueden desarrollar alguna o todas de las funciones listadas a continuación:
- Extracción de datos
- Estandarización de datos
- Integración y consolidación de datos duplicados
- Efectúan reingeniería de los datos y los incorporan en estructuras con arquitectura común.
- Completar datos faltantes, basándose en algoritmos de integridad

- Aplican actualizaciones de los datos, como por ejemplo cambio de la dirección, a partir de una notificación
- Transforman los valores de los datos de un dominio a otro
- Transforman de datos de un tipo a otro
- Calculan datos derivados y sumarizados
- Mejoran los datos, por medio de comparaciones e integraciones con datos de fuentes externas, como la base de datos electoral o de buro de crédito etc.
- Cargan los datos desde la vieja hacia la nueva arquitectura

Tabla 2.4 Proveedores y productos representativos que ofrecen software para la detección de reglas de negocio.

Proveedor	**Nombre del Producto**	**Código de clasificación**
Data Flux Corp.	Data quality Workbench, SmartScrub, Datalogue, Extend, MatchMaker, IntelliMerge, Inspector	CG, A, R, CN
Evoke Software (formeli DB Star)	Migration Arquitect	R
Information Discovery	The Data Mining Suite	R
Innoative Systems	Analyzer, Verify, Dictionary, Edit, Match, Scrub, Household, CorpMatch, Find	CN, CG, A, M, R, P, S
Integral Solutions Ltd Basingstoke, Hampshire, UK	Clemntine	R, CG
MatchWare Tecnilogies (a subsuduary of Vality)	AutoStan, AutoMatch, MatchWare/CL, MatchWare/ PACE	A, R, C
Mobius Inc.	INFOPAC-ABS	A, R
Re-Genisys	Rulefind:R, analize: R	R
Search Software America	NAME 3, EXTENSIONS, Data Cleansing Engine	A, R, CN, P
Trillium Software (a division of Harte-Hanks)	Trillium Software System	CN, CG, R, S, P
Vality Tecnology	Integrity Data Re-Engineering System	CG, R
WizSoft	WizRule, WizWhy	R, A

Limitaciones de las herramientas.

Tanto la reingeniería como la limpieza de los datos no son capaces de corregir todos los datos imprecisos y faltantes. Las mismas limitaciones que existen para las herramientas de análisis se repiten en este caso. Los datos que son limpiados por herramientas automatizadas serán tan limpios como lo permitan las reglas de negocio automatizadas.

Muchas correcciones de errores tendrán que ser eliminados manualmente. Los trabajadores capacitados deben de ser advertidos que la limpieza automática no significa automáticamente precisión. Una valoración física de los datos debe de confirmar el nivel de precisión de los datos limpiados y los datos transformados.

En la tabla 2.5 se presentan los productos que efectúan la reingeniería, limpieza y transformación de datos.

Tabla 2.5 Proveedores y productos representativos que ofrecen software para la limpieza de datos.

Proveedor	**Nombre del producto**	**Código de clasificación**
Ardent Software	Data Stage	CG
Carleton Corp.	Enterprice Integrator, Passport	CG
Century Analisis	CAI Integration Toolset, TDM Interface Engine	C
Constellar Corp.	Warehousebuilder	CG
D2K	Tapestry	CG
Data Flux Corp.	Data quality Workbench, SmartScrub, Datalogue, Extend, MatchMaker, IntelliMerge, Inspector	CG, A, R, CN
DupeKiller, Inc	Dupe Killer	A, CN, S
Evolutionary Technologies	EXTRAXT	CG
GB Information Management	Accelerator, Originator, Enhancer, Address Manager, Postcode Manager	CN, P
Gladstone Computer Sevices	DQ Administrator, Warehouse Quality Administrator	CG, A
Innoative Systems	Analyzer, Verify, Dictionary, Edit, Match, Scrub, Household, CorpMatch, Find	CN, CG, A, M, R, P, S
MatchWare Tecnilogies (a subsuduary of Vality)	AutoStan, AutoMatch, MatchWare/CL, MatchWare/ PACE	A, R, C
Mobius Inc.	INFOPAC-ABS	A, R
OTS Group	Global Thirth Party Name/adress, Cleansing & Enhancment Service	A, CG, CN, SO
Global-Z Intl	Third party name/address cleansing	CN, SO
Group 1 Software Inc.	NADIS: Scrub Master, Search Master, OnLooker, Model 1 Cross-Seller	CN, P

Proveedor	Nombre del producto	Código de clasificación
Hartland	Hartland Warehouse Data Prep System	CN
Hopweiser	Probe, RAINS	CN
Hot Data Inc.	Hot Data, Hot Data Developer Kit	CN, S
Firstlogic Tecnologies	Centric Data Quility Suite	CN, P, S
Information Builder	EDA-SQL, SmartMart, SNAPpack, Data Warehouse	C
Integral Solutions Ltd	Clemntine	R, CG
International Software Products	Thitd-party cleansing	C, SO
Leonard Logic Ltd.	Designer, Scheduler, Engine, Data Links	CG
Master Soft	Nadis: ScrubMaster, SearchMaster, Onlooker, Model MAX	CN, P
Pinnacle Software	Parse-O-Matic	CG
Pitney Bowes Software System	ReUnion	CN, P, S
Platinum Technology	InfoRefiner, InfoPump, InfoTransport, Info-Hub, Decision-Base	CG
Prism Solutions (adquired by Ardent Software, Spring 99)	Prism Quality Manager, Prism Warehouse Executive	C, G, A
Qualitative Marketing Software	Centrus Suite	CN, P
SAS Institute Inc.	SAS Warehouse Administrator, Transformation Engine	CG
SmartDB Corp.	SMART DB Workbanch	C
Search Software America	NAME 3, EXTENSIONS, Data Cleansing Engine	A, R, CN, P
Trillium Software (a division of Harte-Hanks)	Trillium Software System	CN, CG, R, S, P
Vality Tecnology	Integrity Data Re-Engineering System	CG, R

2.2.4 Herramientas para prevenir defectos en la calidad de la información

Las herramientas para la prevención de defectos se usan para automatizar los procesos informáticos. Ellos mejoran la calidad minimizando la introducción de errores dentro de los recursos. Ellos proporcionan automatización de las reglas del negocio, en aplicaciones que crean y actualizan datos. Estas herramientas tienen una variedad de módulos de programas,

subrutinas, librerías de funciones y tablas. El propósito es implementar las ediciones de los datos y la validación de las reglas.

Funcionalidad de las herramientas.

Las herramientas para la prevención de defectos tienen los mismos tipos de funciones como las herramientas de limpieza de datos. No obstante, ellos proporcionan limpieza en línea y no en procesos de tipo lotes (batch).

Tabla 2.6 Proveedores y productos representativos que ofrecen software para la prevención de defectos.

Proveedor	**Nombre del producto**	**Código de clasificación**
DBE Software	DB- Examiner	A, M, P
Gamma Research	OCRProof	A, P
GB Information Management	Accelerator, Originator, Enhancer, Address Manager, Postcode Manager	CN, P
Group 1 Software Inc.	NADIS: Scrub Master, Search Master, OnLooker, Model 1 Cross-Seller	CN, P
Centric Firstlogic Tecnologies	Centric Data Quality Suite	CN, P, S
Innoative Systems	Analyzer, Verify, Dictionary, Edit, Match, Scrub, Household, CorpMatch, Find	CN, CG, A, M, R, P, S
Kismet Analytic Corp.	KisMeta Validator, Analist	P, M
Master Soft	Nadis: ScrubMaster, SearchMaster, Onlooker, Model MAX	CN, P
Pitney Bowes Software System	ReUnion	CN, P, S
QAS Systems	Quick Address, Rapid, Pro, Batch, Names, DataPlus, Address-Point, Updater	C, P
Qualitative Marketing Software	Centrus Suite	CN, P
Search Software America	NAME 3, EXTENSIONS, Data Cleansing Engine	A, R, CN, P
Trillium Software (a division of Harte-Hanks)	Trillium Software System	CN, CG, R, S, P

Limitaciones de las herramientas.

Estas herramientas pueden solo implementar ediciones y automatizar la validación de las pruebas. Proporcionan pruebas razonables, para asegurar la validez de los valores, como son los códigos postales, las direcciones de domicilios, pero no con una certeza absoluta La introducción de información con calidad requiere de suministradores de información entrenados y capacitados. Además el software con el que están trabajando debe de usar catálogos para la validación de los valores que ellos capturan
En la tabla 2 6 se presentan los productos que efectúan la prevención de defectos.

2.2.5 Herramientas para el manejo de los metadatos.

Las herramientas para el manejo de los metadatos proporcionan una administración automatizada y control de la calidad, tanto de la definición de los datos como para el desarrollo de la arquitectura informática.

Funcionalidad de las herramientas.

Las herramientas para el manejo de los metadatos desempeñan una o mas de las funciones listadas a continuación.

- Valida si los nombres de los datos corresponden a los estándares
- Valida las abreviaciones de los nombres de los datos
- Aseguran la existencia de todos los componentes requeridos, para la definición de los datos
- Mantienen los metadatos, para controlar la reingeniería y los procesos de limpieza de datos
- Evaluación de la normalización de los modelos de datos
- Evaluación del diseño de la base de datos, para integridad de llaves primarias y foráneas y optimización del rendimiento.

Limitaciones de las herramientas.

Por medio de las herramientas para el manejo de metadatos se logra la documentación de las especificaciones de los productos informáticos. Los componentes básicos incluyen el nombre del dato, la definición, las reglas de negocio, junto con las relaciones entre los datos de la misma manera como están representados en el modelo de datos o en la base de datos. (Ver tabla 2.7)

Estas herramientas son incapaces de determinar cuando un dato requerido por los trabajadores capacitados falta, o si el dato incluido es definido correctamente.
Las herramientas de la calidad de los metadatos tienen auditorias para asegurarse que los nombres y las abreviaciones cumplen con los estándares, pero no pueden determinar cuando estos estándares son buenos y proporcionan valores útiles para los trabajadores capacitados y cuando no
La siguiente tabla presenta los productos que efectúan el manejo de los metadatos.

Tabla 2.7 Proveedores y productos representativos que ofrecen software para el manejo de los metadatos.

Proveedor	Nombre del producto	Código de clasificación
Compedia	SA Name Cop, EnComp	A, M
DBE Software	DB- Examiner	A, M, P
Innoative Systems	Analyzer, Verify, Dictionary, Edit, Match, Scrub, Household, CorpMatch, Find	CN, CG, A, M, R, P, S
Intellidex (sibsidiario de Sybase)	Warehouse Control Center, User module, Administrator Module, Meta Data Manager	M
Kismet Analytic Corp.	KisMeta Validator, Analist	P, M
Pine Cone Systems	Content Tracker, Refreshment Tracker	A, M

2.2.6 Evaluación de las herramientas para la calidad de la información.

Una herramienta de software se debe de evaluar siempre desde el punto de vista de la efectividad con la que resuelve un problema particular de la empresa y la hace lograr sus objetivos. Lo más importante no es tener la mejor herramienta de todas, si no tener la herramienta más correcta para resolver el problema y alcanzar la calidad deseada. Primero es recomendable definir el problema que se tiene por resolver. Como la información de calidad es un desafío, para cualquier empresa, la evaluación y la selección del software de limpieza tiene que ser iniciativa de la empresa. Después de comprender el problema que se resolverá, se necesita determinar que categoría o categorías de funciones automatizadas de calidad de la información se necesitaran. Por ejemplo, el hecho de tener un data warehouse no siempre significa que el problema tiene que ser resuelto por la limpieza de los datos. El problema puede radicar en que los datos sean defectuosos desde la misma fuente, y por lo tanto los suministradores de la información no saben quien usa la información que ellos crean. Por lo tanto para resolver este problema se tiene que usar la herramienta para la prevención de defectos y se tiene que capacitar el personal que genera la información. Una vez encontrado y definido el problema del negocio, se tiene que trabajar en conjunto con todos los que participan en los procesos involucrados para definir requerimientos para el área de sistemas, para los procedimientos operativos y para los administradores de estos procesos.

El objetivo es traducir estos requerimientos en criterios de evaluación, como lo son:

El *costo del producto, relacionado al valor agregado para el negocio.*
Algunos productos para la calidad de los datos son muy caros, pero por lo general los ahorros que proporcionan pueden ser mucho mayores a lo invertido. El objetivo es resolver fácilmente los problemas de calidad de información, minimizando el costo y maximizando la eficacia de las funciones de calidad desarrolladas.

El precio por tener licencia del producto.
Cuando se trata de adquisición de software el precio generalmente representa sólo la adquisición que es sólo una fracción del precio total. También el costo del producto incluye el costo del mantenimiento y el uso. Es importante mencionar que a veces un producto barato, puede ser más difícil de usar, aprender y el costo de la consultoría y el soporte que se requieren puede ser mayor, que en un producto de mayor precio, pero fácil de usar. Hay productos que son baratos, fáciles de aprender de usar, pero pueden tener menos funcionalidad que otros productos y no siempre son la mejor opción.

Tipos de plataformas soportada por el producto.
Es de suma importancia como el software trabaja en el ambiente informático de la empresa. Si corre sobre el hardware de la empresa o se requiere la adquisición de alguna nueva tecnología.

El soporte para el acceso de datos
Validar si el producto accesa los datos de su ambiente o se tienen que efectuar conversiones permanentes de una plataforma a otra. Lo último eleva el costo y la posibilidad de introducir errores.

Soporte de los tipos de bases de datos/archivo.
Si el producto soporta de manera directa los datos desde las bases de datos y los archivos, o se requiere una conversión de los datos. Esto añade un factor de tiempo y dinero así como aumenta la posibilidad de un error.

Soporte de los tipos de datos.
Evaluar si el producto soporta los tipos de datos que se manejan en la empresa. Ejemplo: El análisis y la limpieza de datos geográficos son irrelevantes si la empresa no maneja los tipos de datos geográficos. Tampoco se requiere un producto sofisticado de limpieza de datos si la empresa tiene solo cincuenta clientes comerciales.

Criterio de identificación de registros duplicados.
Evaluar que tan bien la herramienta utiliza los diferentes criterios de integración para los registros duplicadas.

Criterio de consolidación de registros duplicados.
Evaluar que tan bien la herramienta de limpieza de software soporta las reglas del negocio, para definir y determinar los requerimientos para la consolidación de los datos, como son:

- La prioridad de los archivos.
- Conjuntos de campos dentro de archivos. Por ejemplo, seleccionar los campos uno a tres desde el archivo "A", pero seleccionar campos de cuatro a siete desde el archivo B. El registro de referencia puede agrupar los diferentes campos en archivos diferentes, y no necesariamente todos los campos desde un archivo.
- Criterio de exclusión de registros. Se pueden especificar reglas para cuando seleccionar los mismos atributos desde algunos archivos, como el mas reciente actualizado y el mas viejo actualizado.
- Como se pueden mantener y especificar estas reglas.

Lógica de transformación.
Evaluar que tipo de transformación de datos se proporciona. Si se soportan todos los tipos de datos que existen dentro de la empresa, tanto en la fuente como en el dato resultante.

Facilidad para definir las reglas de transformación.
Validar si la herramienta incluye reglas predefinidas. En caso contrario que tan fácil es su definición y si se requiere su conversión en rutinas o subprogramas

Habilidad para actualizar la fuente y los datos resultantes.
Si el producto de limpieza tiene la capacidad de efectuar la limpieza directo en el recurso o sólo proporciona limpieza para los datos extraídos. Una regla básica recomienda que si el dato se usa desde la fuente, se tienen que corregir ahí mismo. Si el dato es inconsistente entre el fuente y un recurso secundario, las búsquedas en los dos conjuntos de datos llevaran a respuestas inconsistentes.

Múltiples archivos fuentes consolidados en un solo archivo resultante y viceversa.
Si el producto de limpieza proporciona datos desde múltiples archivos para ser procesados y consolidados en un archivo único resultante.

Transformación de múltiples campos.
Si el producto tiene la capacidad de desarrollar transformaciones combinando múltiples campos. Si hay algún limite del numero de campos para que la herramienta que se usa pueda computar los datos derivados y si puede usar lógica Booleana para pruebas y así hacer transformaciones de los datos.

Integración de la herramienta con los repositorios de la empresa.
Evaluar si se tienen que recodificar las reglas del negocio hacia la herramienta o se tienen que extraer las reglas junto con los metadatos desde sus propios repositorios y diccionarios. ¿Puede este repositorio convertirse en el repositorio central, para las reglas? Esto requerirá muchas personas que puedan extraer información desde la herramienta. ¿Existen herramientas de extracción desde las cuales se puedan extraer las reglas del negocio de una manera que puedan integrarse en el repositorio de la empresa?

Flexibilidad para transformar las reglas de negocio.
¿Qué tan fácil es actualizar las reglas de negocio? ¿Se podrán mantener las reglas de negocio centralmente y al mismo tiempo ser validadas desde localidades diferentes? Si las reglas de transformaciones actualizadas puedan ser aplicadas de manera fácil para los datos que fueron transformados previamente por reglas de negocio. Por ejemplo si tiene un conjunto de valores para el dominio, como el valor diez que esta dividido entre los de doce y catorce, basándose en un dato conocido. Si se pueden actualizar automáticamente el dato que tiene como valor diez en un valor doce o catorce que es apropiado.

Precisión y completes de los datos externos.
Si el proveedor esta incluyendo fuentes de datos externos, como códigos postales, datos demográficos, datos de otras organizaciones se debe de validar si la estructura de estos

datos corresponde con la estructura de la empresa, si son completos y precisos. Si se actualizan con frecuencia o tienen un atraso en la actualización

Fáciles de usar.
Como la limpieza de datos y la calidad de la información son actividades costosas, es bueno minimizar el esfuerzo requerido para desarrollarlos. Hay que comparar las herramientas de gran labor, pero comprensivas con las fáciles de usar pero menos comprensivas. Comparar las mejoras de calidad y los tiempos por lo que necesitara soporte técnico.

Capacitación.
Evaluar si se requiere de un entrenamiento de cómo usar la herramienta, que recursos están disponibles y cual es el costo.

Que metodología soporta la herramienta.
Validar si el proveedor tiene metodología para usar la herramienta. Si es así, evaluar si es completa o solo explica las características de la misma. Si es verdadera, y tiene incluidas sus limitaciones. Las empresas por lo general son o proveedores de metodológicas o proveedores de herramientas, las dos funciones simultáneamente se cumplen en muy pocos casos

Soporte para usuarios múltiples.
Evaluar que tan accesible es la herramienta, si requiere de administrador central o puede ser usada por muchas personas en conjunto.

Capacidad para la prevención de datos defectuosos.
Validar si el producto puede ejecutarse en tiempo real desde aplicaciones que crean y actualizan los datos. Validar si el producto usa las mismas reglas de negocio definidas, para la limpieza y validación en los procesos en lotes. Esto incrementa mucho el valor del producto.

Calidad de los resultados.
Validar la calidad de la herramienta por los resultados obtenidos Como corrige y mejora los datos según nuestras expectativas.

2.2.7 Técnicas para la calidad de la información.

Existen cinco principales categorías de técnicas que se usan en el manejo de la calidad de la información Estas son:

Técnica de recolección de información y análisis. Esta técnica nos ayuda a comprender la naturaleza de los problemas de calidad de información.

Técnicas de documentación. Esta técnica se puede usar, para administrar la información y la comunicación de información compartida.

Técnicas de presentación. Se usa para ayudar en la mejora en localidad por medio de comunicación de la valoración de los resultados.

Técnica de mejoramiento y resolución de problemas. Una vez descubiertos los problemas de calidad, esta técnica ayuda para mejorar los procesos o los datos.

Técnicas de control de la calidad. Una vez mejorada la calidad de los datos, esta técnica ayuda a seguir manteniendo la calidad a un alto nivel.

Como técnicas podemos mencionar algunas de las mas usadas ya consideradas como clasicas. Por ejemplo:

- Amplio uso de catalogos para validación
- Creacion de check list
- Análisis de costo-beneficio
- Seguimiento de la satisfacción del cliente por telefono o por cueationario
- Diccionario de datos
- Diagramas de flujo de la información
- Documentos para políticas y procedimientos

En el siguiente punto se describe una de las empresas que crean e implementan todas las técnicas y herramientas mencionadas.

2.3 La empresa "Innovative Systems, Inc"

2.3.1 Perfil de la Empresa

ISI es una empresa establecida en 1968 como pionera en la industria del manejo de información de clientes y continúa siendo el líder mundial en este ramo. Tres décadas de experiencia con diversas organizaciones permiten ofrecer un extenso conocimiento y experiencia adquiridas a través de cerca de 1200 proyectos en 22 países y en 7 diferentes lenguajes.
Innovative Systems, Inc. (ISI) está especializada en la Calidad de Datos en Sistemas de Clientes, es decir la correcta identificación del cliente y de sus relaciones con la institución. cuál es su nombre, cuál es su apellido paterno y cuál es su apellido materno y cuáles son todas las relaciones que tiene con la institución a través de sofisticados procesos de limpieza y estandarización y de identificación de duplicados.

No cabe duda que el propósito fundamental de una base de datos de información de clientes es permitir la exacta identificación de las relaciones entre los clientes y la institución. La exactitud y consistencia de estos datos son factores críticos para asegurar que todas las iniciativas de negocio tales como análisis de rentabilidad, análisis de riesgos, asignación y manejo de límites de crédito, ventas, mercadotecnia y servicios al cliente en general sean manejadas de la manera más efectiva.

ISI ofrece Software y Servicios para la conversión de la información de clientes y para el manejo de la integridad de datos que garantiza lograr y mantener altos niveles de exactitud y calidad en la información.

Esta empresa ha logrado exclusiva experiencia en las características particulares de los datos de clientes en México, los cuales tienen consideraciones especiales tanto en los nombres (dobles nombres y dobles apellidos además de apellidos compuestos), como en las direcciones, donde con frecuencia se tienen estructuras muy complejas. Esta experiencia ha sido adquirida a lo largo de seis años, habiendo logrado construir un Diccionario especial para México con más de 800,000 palabras a través del análisis de más de 80 millones de registros, mismo número que continuará creciendo con la limpieza y el análisis de datos de nuevos clientes mexicanos.

Los sitemas de ISI se ejecutan en plataformas mainframe de IBM y sistemas abiertos como son Windows/NT, Windows 9x y varios UNÍX incluyendo HP-UX y Sun Solaris.

2.3.2 Descripción de los Sistemas

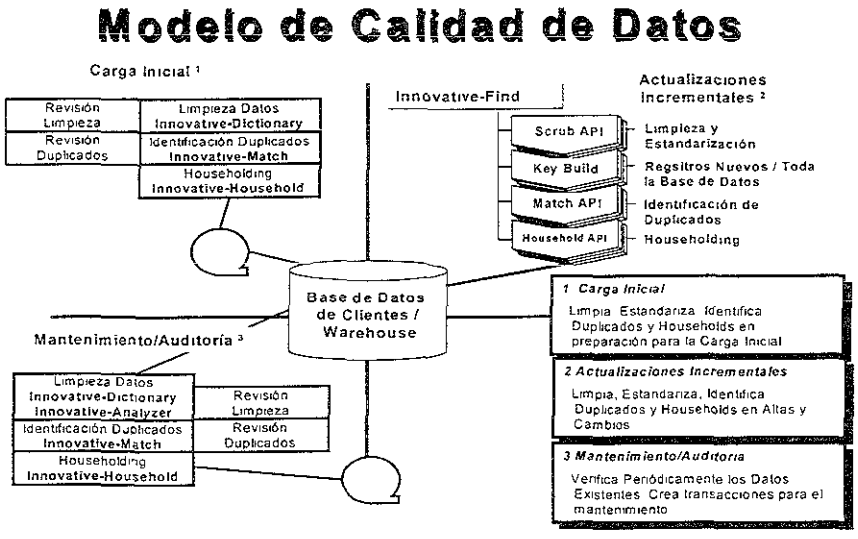

Figura 2.5 Modelo de calidad de datos

En un Modelo de Calidad de Datos podemos distinguir tres etapas fundamentales: (Ver figura 2.5)

1. Carga Inicial En esta etapa se aplica la limpieza, estandarización, identificación de duplicados y households en preparación para la carga inicial del sistema de clientes o un data warehouse.
Las herramientas ofrecidas en esta etapa son:

Innovative-Dictionary	Limpieza y estandarización
Innovative-Match	Identificación de duplicados
Innovative-Household	Identificación de entornos económicos

2 **Actualizaciones Incrementales** En esta etapa al igual que en la etapa uno se efectúa la *limpieza, estandarización, identificación de duplicados y households en altas y cambios.*
La herramienta utilizada en esta etapa es:

Innovative-Find	Ayuda a minimizar el deterioro de la calidad de datos conforme van llegando nuevos registros

3 **Mantenimiento/Auditoría** Verificación periódica de los datos existentes. Creación de transacciones para el mantenimiento
Las herramientas ofrecidas en esta etapa son:

Innovative-Dictionary con Innovative-Analyzer	Análisis, auditoría, limpieza y estandarización
Innovative-Match	Identificación de duplicados
Innovative-Household	Identificación de entornos económicos

En la figura 2.6 se muestran los procesos para la calidad durante la etapa de carga inicial.

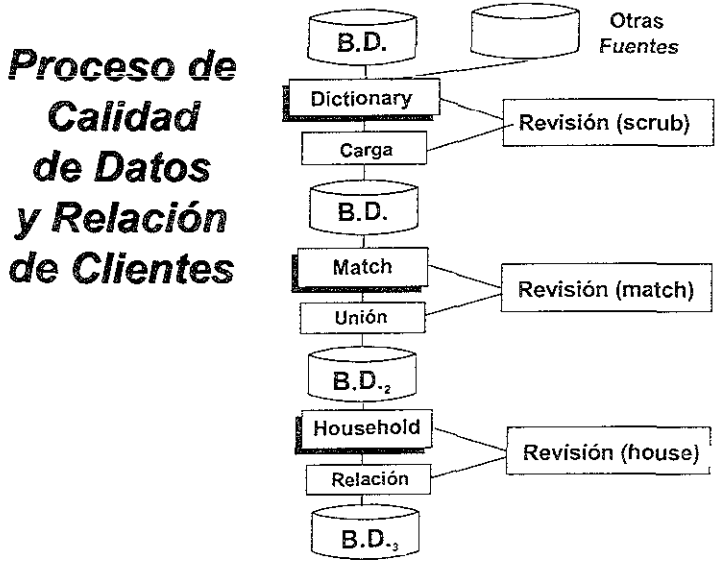

Figura 2.6 Proceso de calidad de datos y relación de clientes.

Innovative-Dictionary System con Innovative-ReviewTool e Innovative-Verify

Innovative-Dictionary es un sistema basado en conocimiento (Knowledge-Based System) para limpieza (scrub) de datos de clientes, el cual utiliza un diccionario de palabras y un diccionario de patrones para comprender las estructuras de nombres y direcciones de los registros fuente, provenientes de cuentas, pólizas, etc., reformatea y estandariza esta información desde el punto de vista de su forma (edit); identifica componentes y también identifica todos los posibles clientes y sus relaciones. Los registros que contienen errores potenciales son segmentados para facilitar su limpieza minimizando así la revisión humana y maximizando la exactitud de los mismos. La versión latino-americana de este sistema incorpora la estructura de dobles apellidos (paterno y materno) y de apellidos compuestos para manejar correctamente las estructuras e identificar mezclas de nombres en el orden apellidos/nombre o nombre/apellidos. El diccionario especial para México se ha venido enriqueciendo con más de 800,000 palabras a través del análisis y la limpieza de más de 80 millones de registros en México. Este sistema también está diseñado para manejar las particularidades de cada país en las estructuras de direcciones.

Innovative-Review Tool proporciona un sistema eficiente para revisión y mantenimiento de los registros marcados como errores potenciales por el sistema de Innovative-Dictionary. Este sistema permite a los revisores el evitar la costosa y tardada tarea del mantenimiento a través de los sistemas en línea de la aplicación del sistema de clientes. A través de su funcionalidad y diseño de la pantalla, los registros pueden ser revisados y corregidos con un en ocho veces de incremento en productividad. El beneficio es una mayor exactitud de los datos como resultado de la limpieza y de una reducción en los costos de revisión, en el tiempo dedicado a esta actividad, así como en los recursos, ya que la revisión se efectúa en un LAN, red local con los archivos en un servidor centralizado.

Innovative-Verify

Innovative-Verify es una componente integral de la licencia de Innovative-Dictionary (no incluido en los servicios externos) que puede ser utilizado para analizar, validar y estandarizar una amplia variedad de tipos de datos que no son nombre y dirección, incluyendo números de teléfono, códigos de producto, números de cuenta o de póliza, números de identificación, etc. Aplicado a varios tipos de información, el sistema Innovative-Verify puede ser utilizado para verificar valores específicos o rangos de valores. Por ejemplo verificar que el campo de edad contenga un número entre 0 y 100, identificar consistencia de datos alfabéticos y numéricos y validar datos contra una lista (comparar código de producto contra una lista de códigos válidos), obtener conteos de frecuencia, identificar recurrencia de problemas, asignar valores por defecto, por ejemplo remplazar datos "basura" como 9999999 por ceros o blancos. Aplicar estándares (estandarizar las diferentes versiones de teléfono – tel, Tel, TEL, Telef, Teléfono, etc.), aplicar

transformaciones simples como justificación a la derecha o a la izquierda, conversión a mayúsculas y minúsculas, remover puntuación o caracteres especiales así como incorporar métodos de validación internos de la institución.

Innovative-Match con Innovative-Review Tool

Innovative-Match es un sistema que incorpora técnicas heurísticas brindando la capacidad de identificar aquellos registros con información de clientes, que pertenecen potencialmente *a la misma persona sin importar las diferencias entre ambos registros atribuibles a errores* de escritura o de ortografía, variaciones en la manera de escribir el nombre o la dirección, o a la utilización de diferentes estándares. Para efectos de identificar los potenciales duplicados, este sistema utiliza hasta nueve elementos obtenidos del nombre y de la dirección, así como RFC´s, Número de Seguro Social u otros elementos determinados por el cliente. Los duplicados potenciales se dividen, basándose en criterios de usuario, en aquellos que pueden ser consolidados automáticamente sin intervención manual y en aquellos que requieren revisión. Este sistema es esencial tanto para la conversión inicial como para los esfuerzos requeridos periódicamente para asegurar una base de datos confiable y libre de duplicados. Este producto es uno de los que más licencias vendidas tiene en el mundo.

La herramienta de productividad Innovative-Review Tool es una herramienta de revisión de duplicados potenciales que incrementa la productividad del proceso de revisión hasta en cuatro veces. Esta herramienta es esencial para disminuir los costos de la limpieza inicial y minimizar también los recursos para la revisión manual asociada con las verificaciones periódicas de la integridad de los datos.

Innovative-Household

El sistema Innovative-Household permite a las instituciones identificar y agrupar aquellos *registros de clientes que pertenezcan al mismo entorno económico* ('household'), para efectos de que aquellas iniciativas alrededor de la información del cliente tales como servicio al cliente, análisis mercadológico, análisis de rentabilidad, análisis de riesgos, etc. no sólo se limiten al conocimiento de un cliente en particular sino que puedan ampliarse también por el conocimiento acerca de otros clientes con los que el cliente está relacionado económicamente. Este sistema establece ligas entre grupos en base a la comparación de elementos de datos que pueden ser tomados del nombre y de la dirección, así como campos definidos por usuario como podrían ser números de teléfono, números de cuenta, etc

Innovative-Household utiliza técnicas de comparación basadas en rangos similares a las de nuestro sistema Innovative-Match.. Cada 'household' se determina como un household aceptable, un household no aceptable o un household cuestionable. Los households cuestionables pueden ser resueltos a través de una herramienta de productividad para revisión llamada Innovative-HouseReview, la cual permite que una persona tome

decisiones lógicas en cuanto a la validez de un household, o en algunos casos a través de una herramienta automatizada de solución de conflictos basada en reglas internas.
El sistema genera llaves de households que pueden ser colocadas en un CIF o base de datos de clientes y pueden ser usadas en la organización para identificar a los clientes y sus households.

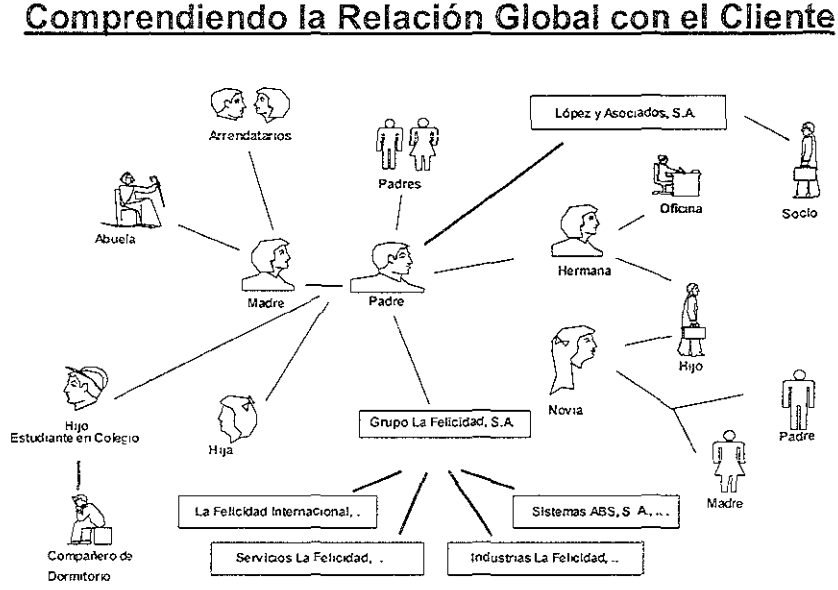

Figura 2.7 La relación global con el cliente.

Por otra parte, como es tradicional en nuestros sistemas, el sistema Innovative-Household proporciona un amplio control por parte del usuario en relación con la manera en la que se desea la identificación de los clientes relacionados. (Ver figura 2.7)
Este sistema está disponible en mainframe de IBM y la herramienta de productividad Innovative-HouseReview se ejecuta en un ambiente de CICS. Se incluye también una herramienta automática para solución de conflictos.

Innovative-Find

Innovative-Find permite a las organizaciones identificar clientes existentes en sus sistemas de información. Este sistema está diseñado para ambientes batch y en línea. Los criterios de búsqueda no se limitan a los datos del nombre y la dirección Esto significa que las organizaciones pueden efectuar búsquedas muy sofisticadas usando nombre, dirección, fecha de nacimiento, RFC, número de teléfono y prácticamente cualquier campo para encontrar rápidamente sus registros de clientes. Innovative-Find crea llaves utilizando estos datos de nombre, dirección, fecha de nacimiento, número de teléfono o cualquier información capturada por una pantalla o una consulta a una base de datos. El sistema permite encontrar a los clientes aún con las diferentes maneras de escribir los datos por errores de ortografía, errores de captura e información faltante. Las llaves son identificadores que son usados por Innovative-Find para encontrar los clientes existentes que son duplicados potenciales y utiliza la misma tecnología de Innovative-Match para clasificar el grado de similitud. Una lista de todos los duplicados potenciales puede ser retornada instantáneamente a la pantalla ordenada según este grado de similitud.

Típicamente, búsquedas en bases de datos pueden resultar en que los usuarios ignoran la búsqueda o la búsqueda falla o los resultados de la búsqueda son muy amplios y los usuarios no utilizan los resultados de la búsqueda, por lo cual se introducen duplicados. (Ver figura 2.8)

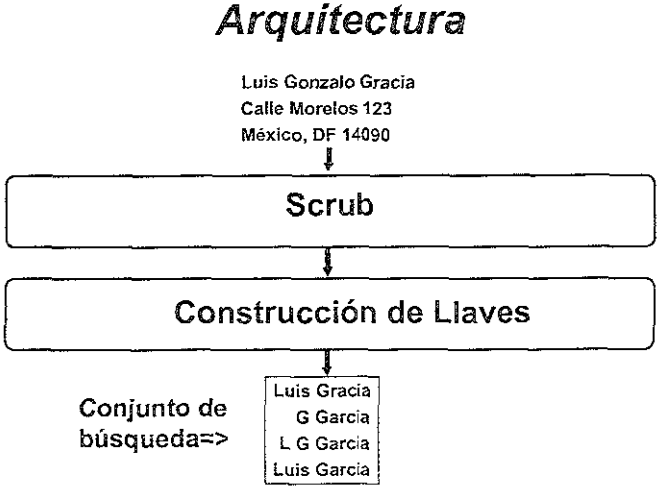

Figura2.8 Etapas para la construcción de llaves de búsqueda para Innovative- Find

La figura 2.8 representa el esquema de funcionamiento de Innovative-Find con las etapas de segmentación del nombre y construcción de llaves para la búsqueda. Las llaves se construyen formando combinaciones del nombre y posibles abreviaciones. Lo que distingue al sistema Innovative-Find es la rapidez y precisión con la cual identifica las posibles relaciones de correspondencias. Este sistema permite, dependiendo de las estrategias elaboradas de búsqueda, escoger cuáles elementos en la información son los más importantes al determinar las relaciones de correspondencia. Otra característica es la edición inmediata en línea, la cual devuelve la información en un formato estandarizado, ayudando a preservar la integridad de los registros nuevos. El sistema evita duplicaciones al encontrar los clientes existentes con precisión. El sistema Innovative-Find es compatible con todos los sistemas mainframe y los sistemas de plataforma abierta.

La arquitectura del sistema representada en la figura 2.9 consiste de tres módulos -- Scrub API para estandarización y construcción de las llaves, Match API para comparación, y sorteo de llaves y Hopusehold API para encontrar relaciones con clientes existentes:

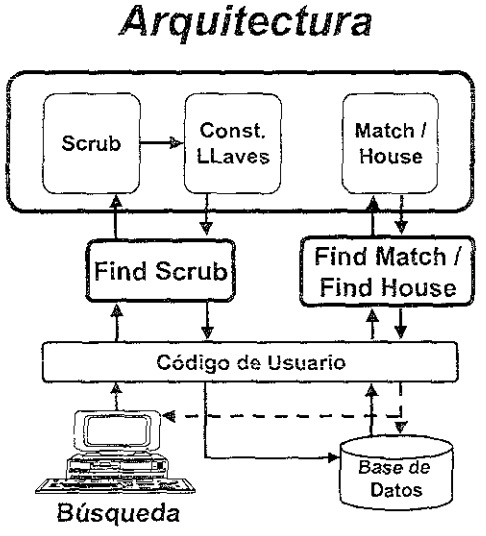

Figura 2.9 Arquitectura global del sistema Innovative- Find

Scrub API
Acepta información de nombres/direcciones en líneas de formato libre o por elementos ya editados con lo que genera llaves y provee consistencia en la información al efectuar la búsqueda. Identifica elementos, estandariza, asigna tipo de línea y registro, lo cual asegura que se han elegido los elementos apropiados para generar las llaves.

Identifica errores para ser verificados por el usuario lo cual controla la calidad de la base de datos. Los errores pueden ser seleccionados para revisión posterior, lo cual provee calidad para el control de auditoría.

Match API
El usuario define el tipo de búsqueda (RFC, Nombre, Dirección, cualquier combinación de éstos) con lo cual las llaves de búsqueda se adecuan a la aplicación. De allí, compara y ordena las llaves usando la tecnología "match" de rangos/criterios. Utiliza tecnología de alias/soundex para generar múltiples llaves para cada tipo de búsqueda y así maximiza las posibilidades de encontrar el registro. Este módulo trabaja con parámetros que permiten al usuario elegir, modificar o agregar llaves, lo que representa una implementación y un ordenamiento de resultados flexible.
El módulo Match API tiene opciones de usuario para almacenar campos de búsqueda con información de Llaves/Indices eliminando la necesidad de índices alternativos en los datos del usuario

Household API
En forma muy similar al Módulo Match API, este módulo encuentra relaciones con clientes existentes.

CAPITULO 3
ALGORITMO PARA INTEGRACION DE CLIENTES

Una empresa decide aplicar una estrategia, para mejorar la calidad de los datos, de la información y el servicio a sus clientes. Además de asegurar la calidad, se necesita asegurar la integridad de los datos y eliminar la duplicidad de la información La estrategia prevé un análisis del software disponible en el mercado para la limpieza de datos, y su adquisición, así como la formación de un equipo de trabajo, para la implementación del software elegido.

3.1. Situación actual y problemática.

3.1.1. Hardware y software utilizado en la empresa.

La empresa tiene una base de datos centralizada de clientes donde se integra toda la información sobre ellos, provenientes de diferentes sistemas y fuentes. No obstante, cada sistema maneja su propio archivo de clientes y prácticamente tiene su propia definición de los datos, de los formatos internos, de los valores, de las reglas de validación y de las reglas de integridad. Por ejemplo: unos sistemas manejan los nombres de sus clientes segmentados por apellidos y nombre, pero no tienen reglas definidas para el orden en que se guardan los apellidos y el nombre. Otros sistemas necesitan grabar el domicilio del cliente, pero no exigen y no almacenan su registro federal de contribuyente (RFC). Debido a estas diferencias existe una gran variedad de los formatos en los que se captura el nombre. Dentro de la empresa existe además de los diferentes sistemas y aplicaciones, también un ambiente de hardware heterogéneo. La empresa tiene un equipo mainframe IBM, varios equipos HP9000 y varios servidores COMPAC. Desde punto de vista del software, la empresa tiene funcionando aplicaciones heredadas, cuyas soluciones se basan en archivos secuenciales o archivos VSAM en el mainframe, otras aplicaciones, en el mismo mainframe, usan el manejador de bases de datos Datacom. Dentro de los equipos UNIX están instalados los manejadores de bases de datos relacionales Sybase SQL Server y Informix. Los equipos COMPAC tienen instalado el SQL Server de Microsoft. Toda esta variedad de hardware y software, junto con las diferentes definiciones de datos y de integridad programadas en las aplicaciones, durante diferentes épocas de su desarrollo, constituyen una falta de estándar único sobre la captura, el almacenamiento, el manejo, la presentación y el uso de la información. En resumen, todas las aplicaciones manejan sus proprios archivos con información sobre los clientes que incluye datos generales y datos demográficos. Muchas veces la estructura de los datos no esta bien definida y representa una mezcla de datos demográficos con datos operativos e históricos. En la mayoría de los casos la información es incompleta y no viene estandarizada. La información se origina en diferentes puntos dentro de la empresa: en las sucursales, ingresa capturada por los clientes en medios magnéticos, o se envía a un departamento para captura centralizada. Esta variedad en el origen de los datos del cliente se muestra en el ejemplo3.1.
Debido que la información de los clientes de la empresa se maneja por cada producto, que ofrece la impresa, el mismo cliente aparece con sus datos generales y demográficos tantas veces cuantos productos tenga contratados con la empresa. En cada repositorio diferente, existe una parte de la información sobre el cliente, según las necesidades del sistema. No hay una imagen integral del cliente. En el caso de los procesos de alta de nuevos clientes, el personal a cargo no siempre obedece las políticas establecidas. Las políticas indican que en el caso de alta de un nuevo producto, lo primero que debe de hacer el empleado es buscar

si el nuevo cliente ya existe dentro de la base de datos de Clientes. En el caso que el cliente existe el nuevo producto debería de asignarse al cliente ya existente. En el caso que el cliente no es encontrado dentro de la base de datos, proceder a darlo de alta.

Ejemplo 3.1:

Los datos capturados en el sistema de cheques son:

NOMBRE, NUMERO	RFC	CALLE	COLONIA	COD. POST.
ACUÑA MARTINEZ MARIA DE LOS ANGELES, 785643	AUMA630401	INSURGENTES 574 B INT 2	ROMA SUR	11560

Los datos capturados en el sistema de depósitos son:

NOMBRE, NUMERO	RFC	CALLE	COLONIA	COD. POST.
MARIA DE LOS ANGELES ACUÑA MARTINEZ, 354466	AUMA630401	INSURGENTES 574 B INT 2	ROMA	11560

Los datos capturados en el sistema de tarjetas de crédito son:

NOMBRE, NUMERO	RFC	CALLE	COLONIA	COD. POST.
ACUNA MARIA DE ANGELES, 898871	AUMA630401 PR2	INSURGENTES 574 B - 2	ROMA SUR	11560

Los datos capturados en el sistema de tarjetas de débito son:

NOMBRE, NUMERO	RFC	CALLE	COLONIA	COD. POST.
MARTINEZ MA DE LOS ANGELES, 4493548	Sin información	INSURGENTES 574 B INT 2	ROMA SUR	Sin informa ción

Los datos capturados en el sistema de acceso por teléfono son:

NOMBRE, NUMERO	RFC	CALLE	COLONIA	COD. POST.
ACU#A MTZ MA DE LOS ANGELES, 8873422	Sin información	INSURGENTES 574 B	ROMA SUR	11000

Como consecuencia el servicio a clientes se ve deteriorado por la falta de una imagen única del cliente. Cuando el cliente se comunica a la empresa, sea para pedir una aclaración o un servicio, el operador que recibe la llamada debe de efectuar una búsqueda en la base de datos centralizada, para validar cierta información sobre el cliente y de esta manera establecer su identidad. Normalmente el cliente se identifica por su nombre y algunas veces por su número de contrato de alguno de sus productos. La búsqueda por nombre puede arrojar varios renglones de datos de clientes homónimos. En este momento el trabajo del operador se dificulta, porque el debe de identificar cuales de todos los renglones en la base de datos se pueden relacionar con este cliente. El operador pierde tiempo analizando la pantalla y esto empeora la calidad del servicio, dado la demora que se produce. A pesar de las dificultades la mayoría de las veces el operador resuelve la situación y la calidad de servicio resultante es relativamente satisfactoria. No así se presenta el caso cuando el cliente quiere consultar el portal de la empresa a través de Internet. Cuando consulta sus productos el no encontrara algunos de ellos lo que provoca un servicio incompleto. Como consecuencia el cliente tiene que hablar por teléfono, para aclarar la situación y esto inconscientemente deja una percepción de mala calidad del servicio. Ahora veremos que sucede cuando el ejecutivo de la cuenta, responsable para atender a este mismo cliente, consulta la base de datos y decide ofrecerle al cliente un producto que no le aparece en la pantalla. No obstante el cliente ya tiene contratado este producto, pero sus datos no se presentan en la pantalla, siendo no integrados por la mala calidad que tiene la información. El ejecutivo habla con el cliente para ofrecerle el producto mencionado y el cliente queda otra vez disgustado pensando que las cosas en la empresa no van tan bien. Por último, dentro de los análisis mensuales que se presentan a los ejecutivos de alto nivel la información sobre este cliente aparecerá con sus datos no consolidados. Estos análisis normalmente sirven para determinar el riesgo de los clientes y tomar decisiones sobre estrategia a seguir para mantener a este cliente y hacerlo crecer con la empresa. Aquí hay dos caminos. Uno, la información se puede quedar incorrecta y de ahí la imagen de este cliente será también incorrecta, y la decisión que se tomara podría ser equivocada. Y dos, el área de sistemas encargada para generar el reporte, debe de efectuar trabajo adicional para consolidar los datos antes de la emisión del reporte. En resumen los nombres y las direcciones de los clientes aparecen en diferentes formatos, sin un estándar único. Falta una imagen integrada sobre el cliente.

3.1.2. Análisis de la calidad actual

El análisis de la calidad que viene a continuación esta hecho según los conceptos y la teoría descritos en él capítulo uno y dos.

Dominio de los datos

El dominio de los datos característicos que manejamos representa los valores válidos que pueden tomar los datos. Los errores de sintaxis existentes en los nombres pueden provocar que estos nombres erróneos sean parte del dominio. Nuestro dominio va a tener valores correctos e incorrectos y la única forma de saber cuales son los correctos y cuales no, es tener un diccionario con todos los nombre y apellidos en México. Tomando en cuenta que hay clientes extranjeros las cosas se complican más. La empresa no dispone de un diccionario de los nombres y apellidos de los clientes. Si un nombre no está bien escrito, es difícil de saber de que persona estamos hablando. Los errores de captura ocurren con

demasiadas frecuencia por lo que representan uno de los objetivos principales de la limpieza de datos que efectuaremos. Tener los dominios bien definidos es uno de los propósitos, para tener calidad de información.

Tabla 3.1. Muestra de la calidad de los datos.

PALABRA	OCCURENCIAS	NUMERO CONSECUTIVO
VELAZQQUEZ	1	123560
VELAZQU	23	123561
VELAZQUE	29	123562
VELAZQUEA	1	123563
VELAZQUELAZQUEZ	1	123564
VELAZQUES	11	123565
VELAZQUESZ	1	123566
VELAZQUEXZ	3	123567
VELAZQUEZ	20002	123568
VELAZQUEZ*	1	123569
VELAZQUEZ-	1	123570
VELAZQUEZA	2	123571
VELAZQUEZALVARADO	1	123572
VELAZQUEZCAMARGO	1	123573
VELAZQUILLO	17	123574
VELAZQUZ	5	123575
VELAZUEUZ	1	123576
VELAZUEZ	3	123577
VELAZUQEZ	1	123578
VELAZWUEZ	2	123579
VELAZZQUEZ	1	123580

El dominio de los datos se establece cuando se separan todas las palabras en la base de datos. Para la separación utilizamos el algoritmo, para la creación del diccionario descrito en este capítulo. Esto nos ayuda a evaluar la calidad del dominio de los datos. En la tabla 3.1. tenemos una pequeña muestra que contiene el apellido "Velazquez". La segunda columna nos muestra la cantidad de ocurrencias de cada palabra y vemos que hay 20002 ocurrencias donde el apellido esta bien escrito, pero también vemos que hay una cantidad de casos donde se cometieron los siguientes errores:

- "VELAZQUEZALVARADO" El apellido y el nombre no están separados con espacio.
- "VELAZQUESZ" Error de captura
- "VELAZQUEZ*" Error de captura
- "VELAZUQEZ" Error de captura

Es obvio que ningún algoritmo de búsqueda de apellidos podrá encontrar en este caso alguna similitud. El análisis en la base de datos muestra que la limpieza no es suficiente y que todavía se encuentran errores. De 20108 ocurrencias del apellido, 106 son erróneos, lo que representa un porcentaje de error de 0.527%. Esto es solo un ejercicio, pero si se quiere

efectuar una evaluación de la limpieza en la base de datos, se tiene que usar una muestra con el diez por ciento del contenido total.

Integridad de los valores de los datos

La integridad de los valores de los datos va estrechamente ligada con el dominio, porque si el dominio esta bien definido también la integridad de los valores de los datos se puede determinar bien. En nuestro caso vimos que el dominio no se puede definir bien en todos los casos. La única manera de lograr esto es limpiando sistemáticamente los datos de los clientes.

Tabla 3.2. Ejemplos de defectos en la integridad.

Numero de cliente	Nombre	RFC	Cod. Postal	Dirección	Estado
64903578	X GARCIA JOSE MANUEL	GAXJ520119	34000	PRIVADA LAGUNA NO 218 COL. INSURGENTES	DURANGO
64906225	SOTO CHAVEZ MAURA	SOCM571121	00000	INDEPENDENCIA NO. 9 SUR SAN FELIPE	TOLUCA
64938731	GOMEZ LAGUNES CARLA	GOLC700718	91940	PASEO FLORESTA SUR NO. 43 FRACC. FLORESTA	VERACRUZ
64938731	GOMEZ LAGUNES CAROLINA	GOLC700718	91940	PASEO FLORESTA SUR NO. 43 FRACC. FLORESTA	VERACRUZ

En la tabla 3.2. tenemos varios ejemplos de mala integridad de los valores. En el primer caso falta uno de los apellidos. El operador capturo una "X". Observamos que la RFC también tiene una "X" en la posición del segundo apellido. El código postal en el segundo ejemplo tiene el valor de cinco ceros. Este valor si satisface la regla que determina que el código postal debe de tener cinco dígitos. Pero el número 00000 no es un código postal real. En esta situación se deben de recomendar dos medidas. La primera que los valores del código postal se deben de validar contra el catálogo de SEPOMEX (Servicio Postal Mexicano). La segunda medida indica que se debe de efectuar un proceso de limpieza que detecta todos los valores de cogidos postales fuera del catalogo y que estos se corrijan manualmente. Una medida muy real de la calidad del código postal es la cantidad de correo devuelto. En la línea tres y cuatro tenemos un ejemplo de RFC incompleto, porque falta la homoclave. En la mayoría de los casos la falta de homoclave no nos afecta, porque la combinación de nombre y fechas de nacimiento se repiten con muy poca probabilidad. En este caso hablamos de dos gemelas, ya que su fecha de nacimiento es la misma, viven en le mismo lugar y se apellidan igual. Si algún departamento efectúa conciliaciones por RFC se puede equivocar por la falta de homoclave. Nuestro algoritmo no se debe de equivocar, porque debe de detectar la diferencia entre los nombres.

Integridad de la estructura de los datos

Este subconjunto de la integridad de los datos, especifica las relaciones existentes entre los datos. En nuestro caso esto significa que el orden en que se escriben los apellidos y el nombre de una persona tiene que especificarse y respetarse. Para el propósito de este trabajo se considera que existe una integridad de la estructura de los datos cuando el nombre que identifica una persona es escrito en la siguiente orden: Apellido paterno,

Apellido materno, Nombre. Cualquier otro tipo de orden provocara una mala calidad en la estructuración de los datos y por lo tanto estos datos serán objeto de limpieza.

Integridad de los datos derivados

El recurso sobre el cual se efectuará la limpieza a primera vista podría contener una gran cantidad de datos derivados. Un ejemplo es la relación que existe entre el RFC, el nombre y la fecha de nacimiento de una persona. Los datos que conforman el RFC son datos derivados de los datos antes mencionados de esta persona. El RFC tiene que ser capturado de manera correcta, para que se tengan datos limpios y bien integrados. La única validación que hace la empresa en este caso es que la regla de los valores de que especifica que las primeras cuatro posiciones son alfabéticas y las siguientes seis numéricas. Esta validación según esta regla resulta insuficiente, porque los trabajadores irresponsables que originan la información en muchos casos teclean el valor "AAAA000000". Este valor si es válido según la regla, pero no es real. Otro ejemplo es la relación entre el código postal y la colonia de la dirección del cliente. El catalogo de SEPOMEX contiene esta relación y se puede usar para validaciones. Aunque los dos ejemplos parecen ser datos derivados, en realidad no lo son, ya que en los valores se capturan por separado y no existe un algoritmo que genere un dato a partir de otro. Más bien se trata de datos relacionados o dependientes. Por lo tanto, para el propósito del análisis estos datos se tienen que considerar como datos independientes.

Datos redundantes

Los recursos con los que se dispone tienen una cantidad considerable de datos redundantes. Esto sucede, porque las características de un dato se repiten en uno o más lugares. Muchos clientes figuran varias veces dentro de la base de datos. En nuestro caso el análisis muestra que la redundancia se debe a errores en la operación. Para poder asegurar la calidad de la información, tenemos que identificarlos y mantenerlos integrados de manera apropiada. Si la limpieza logra evitar estas redundancias la calidad de los datos aumentara de manera significativa y se cumplirá uno de los objetivos establecidos de tener una imagen única e integrada del cliente.

Completes

La completes es un indicador que determina si los datos son suficientes, para poder cumplir con la demanda de información. No existe completes cuando algunos de los datos de un cliente faltan. Para nuestro algoritmo la completes abarca el nombre, el RFC y dirección. La completes del nombre es una condición indispensable. Sin la existencia del nombre completo ningún algoritmo puede funcionar correctamente. La completes del RFC es también obligatoria, para detectar si hay igualdad entre personas. Hay muy pocas ocasiones cuando la falta de RFC se puede reemplazar por la información de los productos relacionados, que puede tener el mismo cliente. La dirección se usa como un criterio secundario de comparación. Por esto, su completes es deseable, pero no obligatoria. Lo que si es importante es la existencia de un código postal valido. El análisis detectó que la falta de completes ocurre con el RFC y el código postal de la dirección. Tener una mejor completes es uno de los objetivos principales de la calidad de información a la que se quiere llegar. Si falta uno de los dos, el RFC o el código postal todavía se puede establecer, mediante revisión manual la identidad del cliente. Pero si faltan los dos atributos al mismo tiempo ya no habrá forma de encontrar al cliente. La completes de los datos influye sobre la

lógica del algoritmo y facilita la decisión de integración automática. Lo que significa que los datos son completos y iguales el algoritmo de manera automática integra los dos clientes. La falta de completes obliga al algoritmo a mandar a los datos para validación manual o los rechaza. La tabla 3.3 muestra dos registros que tienen la suficiente completes de datos para la integración de los dos clientes. La única diferencia es que en el segundo renglón el nombre aparece capturado en el orden incorrecto. Aun que en este caso no existe la integridad de la estructura del dato, vemos que la completes existe.

Tabla 3.3 Ejemplo de completes suficiente en la integridad.

Numero de cliente	Nombre	RFC	Cod. Postal	Dirección	Estado
64922529	CARLOS MAURICIO TUEME PEDRAZA	TUPC661104RA5	83145	CERRADA DEL VIGIA 35 PUEBLO ALTO	SONORA, HE
27719988	TUEME PEDRAZA CARLOS MAURICIO	TUPC661104RA5	83145	CERRADA DEL VIGIA 35 PUEBLO ALTO	SONORA, HE

Tabla 3.4 Ejemplo de insuficiente completes.

Numero de cliente	Nombre	RFC	Cod. Postal	Dirección	Estado
23166176	GARCIA BORJA CARLOS A	GABC671028	55390	AMECAMECA NO 20 ALTAVILLA ECATEPEC	EDO MEX
20094694	GARCIA BORJA CARLOS ALBERTO		00000		

La Tabla 3.4 tiene ejemplos de insuficiente completes. Aunque el primer registro tiene los datos completos, el segundo registro tiene varios datos faltantes. La completes no se tiene que buscar solo en registros separados, sino más bien en grupos de registros, esto es un requisito para poder efectuar la integración entre los clientes.

3.2. Necesidades de la empresa

La empresa comenzó un cambio de enfoque en su política. Anteriormente el enfoque estaba orientado hacia los productos, hoy el día la necesidad del negocio y la creciente competencia del mercado requiere un enfoque hacia el cliente. Se requiere conocer al cliente con todos sus productos integrados. De esta manera la imagen del cliente se vuelve completa y se pueden evaluar el costo del cliente, las ganancias que el cliente aporta a la empresa, calcular y estimar el riesgo del cliente, conocer su lealtad etc. En este sentido la calidad de los datos en la Base de Datos de clientes depende de la integración de los clientes existentes. El objetivo es tener cada cliente con su número único en la base de datos integrada y de esta manera tener su perfil completo. Toda la información de clientes se esta acumulando en una base de datos única llamada CIF (Customer Information File). Esta base de datos radica en un SQL Server de Sybase en un equipo UNIX.

La alta de clientes dentro del CIF se efectúa en modo batch y en línea. La información de los nuevos clientes que fueron dados de alta en los aplicativos se integra en el CIF vía diferentes interfaces. Las interfaces corren como procesos batch de carga de datos nuevos. Las actualizaciones y las correcciones de los datos generales y demográficos de cada cliente se pueden efectuar en dos maneras, uno en línea a través de front end manejado por áreas que se ocupan a mantener la calidad de la información y dos, por procesos batch donde se incorporan todos los cambios efectuados en los sistemas. A pesar de las existentes vías de actualización y limpieza de datos, la información de cada cliente puede permanecer mucho tiempo repetida y sin unificar.

Para mantener la integridad de los clientes se utiliza la herramienta de Innovative systems Inc, llamada ISI-MATCH (descrita en el capítulo 2). El proceso de integración de clientes se efectúa en varios pasos. Primero se extrae la información completa de cada cliente. La información necesaria esta compuesta por los siguientes campos: nombre, apellido paterno, apellido materno, registro federal de contribuyentes y la dirección del cliente. La dirección incluye el nombre de la calle, código postal, nombre de la colonia, nombre de la delegación o municipio y estado de la república. En el segundo paso el archivo extraído se transfiere al mainframe IBM. En el mainframe esta instalado el software de Innovative Systems Inc, ISI-MATCH. El tercer paso es correr el proceso de match y obtener los archivos de salida correspondientes. El proceso de match maneja rangos de posible integración de dos o más personas. Al final del proceso cada pareja de registros se graba junto con su rango determinado por la corrida. Para determinar el rango, el software de ISI-MATCH actúa en varios pasos: primero segmenta el nombre o con otras palabras determina cuales palabras representan el nombre de la persona, cuales determinan su apellido paterno y cuales determinan su apellido materno. Para este propósito se están usando diccionarios internos donde hay información sobre todos los posibles nombres y apellidos en español. Existe otro diccionario de las abreviaciones que se usan para los nombres y los apellidos. Por ejemplo, MARIA EUGENIA y MA EUGENIA, el apellido RODRIGUEZ y RGEZ.

Una vez segmentado el nombre, es fácil comparar los nombres y los apellidos de dos personas y determinar si son iguales.
Después de determinar que los nombres son iguales, si existen, se comparan los registros federales de contribuyentes de las dos personas.
Al final se compara la dirección para determinar si las direcciones de las dos personas son iguales. Para esta comparación se usan los códigos postales y el texto de la calle.

Según el resultado de la comparación de todos los componentes (nombres, apellido, rfc, códigos postales, nombres de las calles), cada pareja obtiene el rengo de igualdad y coincidencia que tiene en estos datos. Según las necesidades de la empresa sobre la calidad de los datos, se deben de determinar que rangos representan interés para ser procesados a continuación. Por ejemplo la empresa puede determinar que los rangos de 1 a 17 según las definiciones de necesidad de la calidad de la información, representan una certeza de 100 % que las dos personas son idénticas. Los rangos de 200 a 999 representas demasiadas diferencias para que las dos personas sean idénticas. Mas bien - se trata de personas totalmente diferentes.

Los rangos de 17 a 200 representan una gran probabilidad que las dos personas sean idénticas pero los datos existentes son incompletos y no coinciden a un 100%. Los registros calificados en estos rangos necesitan una revisión manual por personal calificado

El siguiente paso en el proceso de match es extraer los registros de parejas de clientes con rangos de 1 a 17 y transferirlos de regreso a la plataforma UNIX. Después un proceso en UNIX aplica la integración de cada pareja de clientes dentro de CIF y actualiza todos los datos necesarios. Cabe mencionar que si un cliente aparece mas de dos veces se forman varias parejas para la integración de tal manera que al final debe de quedar un solo cliente.

En otro paso se extraen los registros de rangos 17 a 200 y se transfieren a un servidor COMPAC como un archivo de trabajo. En este momento el personal calificado empieza a trabajar sobre la validación manual de las posibles parejas iguales. El personal usa otra herramienta de Innovative Systems Inc, llamada PC-REVIEW. Con al ayuda de la herramienta el personal visualiza en la pantalla los datos de los dos clientes y los compara visualmente o efectúa las verificaciones pertinentes de la documentación existente. Después se toma la decisión si la pareja se debe de integrar o no, conforme existen los datos suficientes para su integración. Las parejas no apropiadas para la integración se ignoran. Las parejas por integrarse se marcan por el software y después se extraen en un archivo y se procesan.
Como se observa de la descripción del proceso, este último resulta un gran consumidor de tiempo para la extracción de los datos y la transferencia entre plataformas heterogéneas. Más todavía tiempo se necesita para la parte de la revisión manual. Según la cantidad de los registros por revisar y según la complejidad de la revisión y según la cantidad de las personas que revisan, un proceso puede durar de dos días hasta dos semanas.

Considerando todo estos factores y otros internos de la empresa como ocupación del personal a cargo, resulta que en realidad entre dos procesos de integración puede pasar más de un mes. En la vida real pasan de tres a cuatro meses. Con otras palabras en el periodo de un año se corren de tres a cuatro integraciones con el software de ISI-MATCH. Esto es completamente insuficiente para responder a las necesidades reales de la empresa para calidad de la información. Se necesita más oportunidad para poder detectar clientes iguales pero no integrados. Innovative Systems Inc ofrece otra herramienta, para la integración diaria, pero esta tiene un costo elevado. Por esto la empresa decidió desarrollar su propia herramienta, para efectuar diario integraciones de los nuevos clientes. Esta herramienta debe de solucionar únicamente el problema de los nuevos clientes. Para los clientes ya existentes se seguirá aplicando el proceso de ISI-MATCH.

3.3 Propuesta para solución del problema.

La idea conceptual de la propuesta es tener definido el dominio de todas las palabras usadas en los nombres. Para el propósito se crea un diccionario con todas las palabras ubicadas en los nombres de los clientes. A cada palabra encontrada se le asigna un numero único (normalización de las palabras). La solución debería de buscar el nombre del cliente nuevo dentro de este dominio Este dominio se puede llamar un repositorio de todas las palabras contenidas o se puede llamar un diccionario. Junto con la creación del diccionario se debería de tener la relación entre cada palabra del diccionario y el numero de cliente

donde esta se encuentra. Esta relación se va a llamar referencia cruzada. En resumen vamos a crear dos tablas, la primera es del diccionario y la otra es la referencia cruzada.
La tabla 3.5 presenta un ejemplo de extracción del diccionario. En la primera columna se encuentra la palabra, en la segunda las veces que esta palabra aparece dentro de toda la base de datos, y la tercera contiene un identificador numérico único de la palabra. Este identificador sirve para efectuar la liga con la tabla de las referencias cruzadas.

Tabla 3.5 Formato del diccionario

PALABRA	VECES	ID CONSECUTIVO
ACUNA	525	4250
ACURI	1	4251
ACUSTICA	1	4252
ACUÑ	1	4253
ACUÑA	1670	4254
ACVEDO	1	4255
ANGELES	14186	9944
ANGELICA	11285	9963
MARI	440	79119
MARIA	203167	79120
MARTTELO	6	80060
MARTTINEZ	5	80061
MARTUSCELLI	1	80062
MARTY	13	80063
MTZ	14	80065

En la tabla 3.6 se presenta la estructura de la referencia cruzada. La primera columna es el identificador numérico único de la palabra y sirve como llave foránea para la liga con la tabla del diccionario. La segunda columna indica la posición de la palabra dentro del nombre. La tercera columna indica el numero del cliente cuyo nombre contiene esta palabra.
Utilizando el diccionario y la referencia cruzada, podemos buscar las palabras del nombre de un nuevo cliente y encontrar los números de clientes ya existentes donde se encuentran estas mismas palabras. De esta manera detectamos los clientes existentes, los cuales son homónimos al nuevo cliente.
El diccionario contiene todos los valores de palabras existentes dentro de los nombres en la base de datos. Esto incluye los valores correctos (las palabras escritas de manera correcta y sin errores de ortografía) y también incluye los valores fuera de rango como son las palabras escritas con errores (palabras pegadas, palabras escritas con acentos, palabras abreviadas, palabras con falta de letras etc). En este sentido la importancia del diccionario es todavía mayor, porque este se puede utilizar para otros tipos de actividades de limpieza.

Tabla 3.6 Formato de las referencias cruzadas.

ID CONSECUTIVO	ORDEN	NUMERO DE CLIENTE
4250 (ACUNA)	1	234599
4254 (ACUÑA)	1	785643
4254 (ACUÑA)	3	354466
9944 (ANGELES)	4	354466
9944 (ANGELES)	4	898871
9944 (ANGELES)	5	4493548
9944 (ANGELES)	6	8873422
9944 (ANGELES)	6	785643
79120 (MARIA)	3	785643
79120 (MARIA)	1	354466
79120 (MARIA)	2	898871
80061 (MARTTINEZ)	2	785643
80061 (MARTTINEZ)	6	354466
80061 (MARTTINEZ)	1	4493548
80063 (MARTY)	2	987623
80065 (MTZ)	2	8873422

Por ejemplo se pueden detectar palabras con errores de ortografía, palabras abreviadas o palabras mal escritas y fuera del estándar. El diccionario viene ordenado por orden alfabético y entonces si el nombre "Ordoñez" esta mal escrito con una letra "n" en lugar de "ñ", en el diccionario aparecerán dos palabras una Ordonez y otra Ordoñez , esto significa que existe algún cliente cuyo nombre esta mal escrito. Organizado de esta manera, el diccionario es una herramienta poderosa para poder ubicar de manera rápida y eficaz cualquier palabra que necesitamos encontrar sin importancia si esta aparece al principio de un nombre o en el medio o al final. Como ejemplo podemos
utilizar la búsqueda de nombres que contienen la letra Ñ pero que esta aparece capturada con el símbolo de ampersand (&) . Para este propósito es suficiente efectuar una búsqueda del símbolo "&" dentro del diccionario. La búsqueda es muy rápida debido a que el diccionario contiene alrededor de 100000 palabras. Una vez detectadas las palabras que contienen el símbolo "&", utilizamos su numero de identificación numérico único, para usarlo como llave y buscar en la tabla de las referencias cruzadas todos los números de registros donde se encuentran las palabras detectadas. Esta búsqueda también es muy rápida debido a que la referencia cruzada tiene un índice organizado por el identificador numérico único de palabra.

La figura 3.1 representa el flujo de la información dentro del proceso de integración diaria de nuevos clientes. El esquema muestra las cuatro fases del proceso. La primera fase es el alta diaria de nuevos clientes con sus productos en la base de datos de clientes. La segunda fase consta de los procesos de creación y actualización del diccionario. Según el volumen de información ingresada se puede efectuar diario o semanalmente. La tercera fase esta representada por el algoritmo de extracción e integración de los nuevos clientes. Una vez concluido el proceso de integración se evalúan los resultados y la información puede pasar por dos caminos. Cuando una pareja de clientes es reconocida de manera automática se

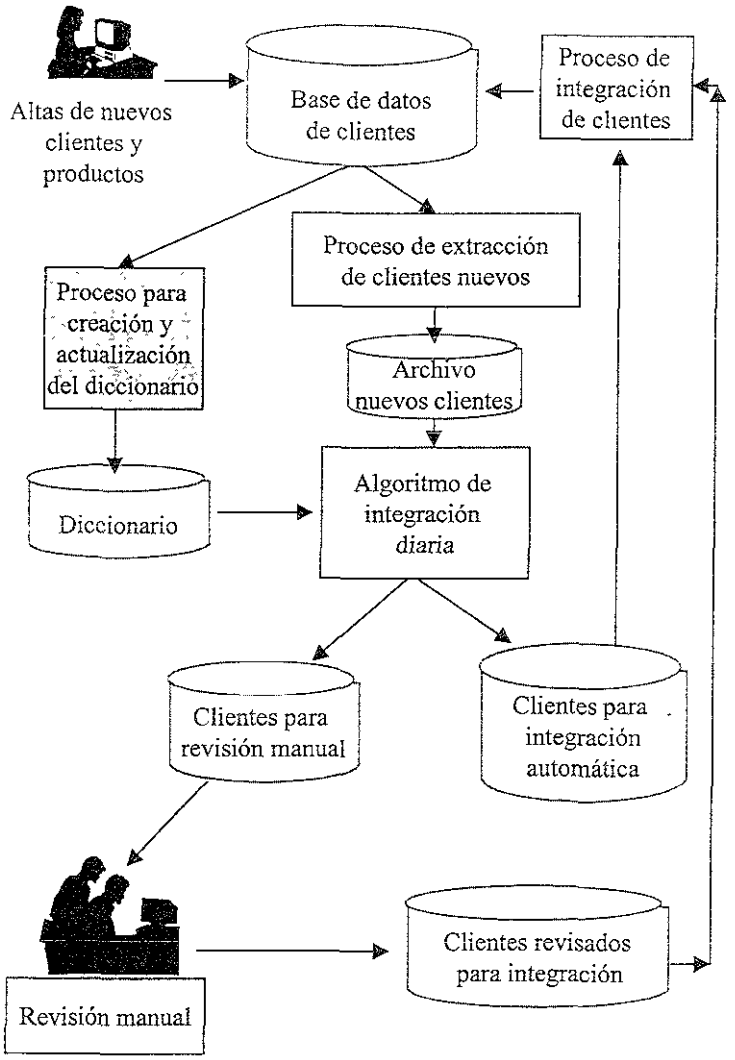

Figura 3.1 Proceso de integración diaria.

pasa directo a proceso de integración de clientes con lo cual termina el ciclo. Si la pareja se manda para una revisión manual la información entra en la fase cuatro donde el personal capacitado hace revisión por pantalla y determina si la pareja se rechaza o se manda para integración.

3.4 Algoritmos.

3 4.1. Descripción del algoritmo para la creación del diccionario

Para el proposito del algoritmo vamos a utilizar los siguientes campos de la base de datos de clientes.
Número de cliente
Apellido paterno
Apellido materno
Nombre
RFC
Tipo de cliente – persona física o moral

Los pasos del algoritmo se describen a continuación· (Ver figura 3.2)

Paso 1.
Extraemos en tres archivos diferentes todos los nombres. En el primero los apellidos paternos, en el segundo los apellidos maternos y en el tercero los nombres. Los apellidos y los nombres se separan en tres archivos, porque necesitamos marcar en la tabla de las referencias cruzadas el lugar donde fue encontrada cada palabra, si fue en el paterno, materno o en el nombre. El análisis previo determinó que las palabras pueden ser separadas, por blancos, paréntesis, puntos, comas, diagonal, y dos puntos. Debido a que en un apellido paterno o materno puede existir mas de una palabra, se usan estos símbolos como separadores de palabras. Junto a cada palabra viene él numero de registro de cada cliente donde fue encontrado. Esta información nos va a servir en el paso cuatro.
Al finalizar el paso ya tenemos todas las palabras encontradas en la base de datos, pero repetidas varias veces.

Paso 2.
Se aplica un sort con la opción "unico" sobre los tres archivos concatenados para eliminar las repeticiones. De esta manera obtenemos el diccionario en el cual cada palabra aparece una sola vez.

Paso 3.
En este paso se le asigna un número consecutivo único a cada palabra y ya tenemos el diccionario hecho. Esto se hace porque los sistemas trabajan mejor con variables numéricas que con variables tipo texto.

Paso 4.
Con base al nuevo diccionario obtenido y los archivos de paso uno, se crea la tabla de las referencias cruzadas, donde se relaciona cada numero de cliente con los números de palabras que aparecen dentro del nombre de este cliente.

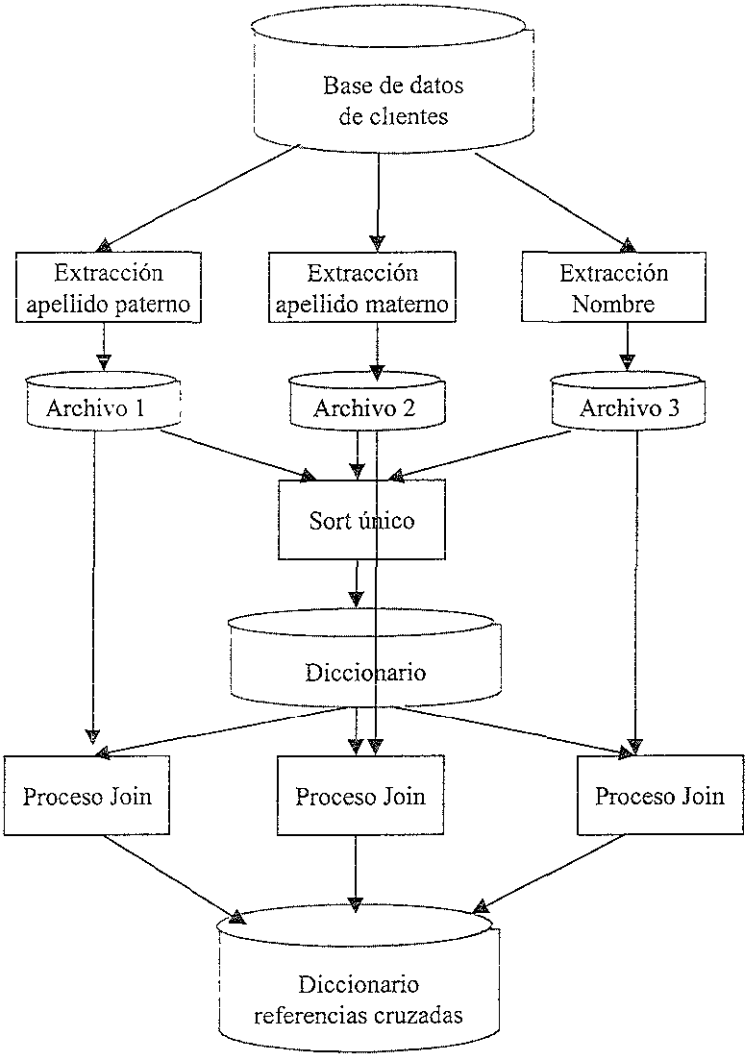

Figura 3.2 Creación del diccionario.

Paso 5.
Al final con base del nuevo diccionario y los archivos extraídos del paso numero uno se calcula él numero de ocurrencias de cada palabra dentro de toda la base de datos de "clientes". Con este numero actualizamos la columna dos del diccionario. Este numero nos servirá para la optimización de los procesos de búsquedas posteriores. Este numero de ocurrencias se utiliza para optimizar el tiempo de la búsqueda. Supongamos que buscamos un cliente cuyo nombre tiene tres palabras A, B, C. Supongamos que la palabra A tiene 1000 ocurrencias en la base de datos, la B tiene 5 ocurrencias y la C tiene 30 ocurrencias. Si empezamos la búsqueda por la palabra A, que tiene 1000 ocurrencias, tendríamos que rastrear 1000 registros para buscar adentro las palabras B y C. De lo contrario si empezamos la búsqueda por la palabra B vamos a encontrar 5 ocurrencias y utilizaremos menos tiempo buscando las palabras A y C dentro de los 5 registros donde ya encontramos la palabra B.

3.4.2. Descripción del algoritmo para búsqueda de homónimos

Cada día dentro de la base de datos de "clientes" se insertan nuevos registros. Estos provienen de los diferentes sistemas de la empresa. Es muy probable que algunos de estos "nuevos" clientes, sean clientes ya existentes, pero dentro del sistema aparecen como nuevos, por algún error humano o por cualquier otra razón. Este algoritmo tiene dos fases principales. Durante la primera fase para cada uno de estos clientes se buscaran sus homónimos dentro de la base de datos, utilizando el diccionario y las referencias cruzadas, formando parejas de clientes con los nombres iguales. Posteriormente en una segunda fase, según la información disponible, se determinará si los clientes de cada pareja son la misma persona o nada mas se llaman igual y son diferentes personas.
El algoritmo tiene como entrada dos campos, el número del nuevo cliente y su nombre ubicado en un solo campo, es decir, no tenemos separados el apellido paterno, el apellido materno o el nombre, por la razón que no sabemos como estaba capturada la información. *Normalmente los sistemas no tienen una regla estricta que determine el orden de captura de* los nombres. Si el orden va a ser Apellido paterno, Apellido materno, nombre o nombre Apellido paterno y apellido materno

El análisis previo sobre la captura de datos detecto que para representar la letra "Ñ", se usan varios símbolos. La mayoría de estos son combinaciones en código hexadecimal. En otros casos se usa el símbolo de numero o ampersant. (Ver figura 3.3)

Paso 1:
Unificación de todos los símbolos que se utilizan para representar la "Ñ". Estos son ampersant, arroba, número y todos los símbolos detectados. Después de ubicar el símbolo este se convierte en código decimal "D1" que es la representación de la "Ñ" en UNIX.

Paso 2:
Este paso trata de la conversión de todos los separadores de palabras posibles en uno solo que para mayor conveniencia será el de "blanco". Los separadores que pueden ser usados en lugar de "blanco" son: parentesis, diagonal, coma, punto, punto y coma, y dos puntos.

Después de la conversión se eliminan también los dobles espacios y se convierten en un solo espacio.

Paso 3:
Se separan las palabras y se cargan en una matriz temporal de trabajo MATRIZ1. La matriz guarda también el orden de las palabras. Cual es la primera, cual la segunda y cual la tercera. También se memoriza la cantidad total de palabras. Guardamos el nombre completo en una variable NOMCOMPLETO, separado por blancos, para comparación en el paso 11.

Paso 4:
En el caso de que existan nombres de menos de tres palabras no se deben de considerar, porque los datos son incompletos y no son suficientes, para la comparación.

Paso 5:
Se determina si alguna de las palabras aparece mas de una vez. Ejemplo: En el nombre Alvaro Pérez Pérez, la palabra Perez aparece mas de una vez. Este indicador se va a usar mas tarde en la comparación de los nombres.

Paso 6:
De la matriz temporal de trabajo MATRIZ1, generada en el paso tres, se escoge la palabra con menos ocurrencias en el diccionario (el segundo campo del diccionario). Esto se hace porque permite la reducción del tiempo y limita el número de iteraciones, que se deben de realizar en las siguientes busquedas. Si por ejemplo buscamos Heréndira Gomez Rodriguez y empezamos a buscar por Gómez tendríamos 100000 posibilidades, en cambio con el nombre de Heréndira tendríamos quizá solo 180.

Paso 7:
De la tabla de las referencias cruzadas extraemos los registros de los clientes que contienen la palabra con menos ocurrencias y obtenemos una tabla temporal de trabajo TEMP1, como un subconjunto de los números de clientes posibles ya muy limitados. La búsqueda se hace por el primer campo que es el identificador numérico único de la palabra.

Paso 8:
Para el resto de las palabras de la MATRIZ1, efectuamos un ciclo buscando cada palabra dentro de la tabla TEMP1 y al final de este ciclo estamos obteniendo todos los clientes que contienen por lo menos una de las palabras del nombre buscado.

Paso 9:
En este paso se determina en que numero de cliente se encontraron al mismo tiempo todas las palabras que estamos buscando, pueden ser 3, 4 o 5.Ahora tenemos que determinar el cliente que contiene todas las palabras a la vez. Esto es muy fácil, porque el numero de cliente que contiene todas las palabras aparecerá dentro del subset tantas veces como cuantas palabras estamos buscando.Este paso tiene dos variantes. La primera variante es cuando todas las palabras del nombre nuevo son diferentes y la segunda es cuando una de las palabras aparece mas de una vez. (al caso del paso cinco de Alvaro Pérez Pérez). El resultado de la búsqueda se graba en la tabla temporal TEMP2.

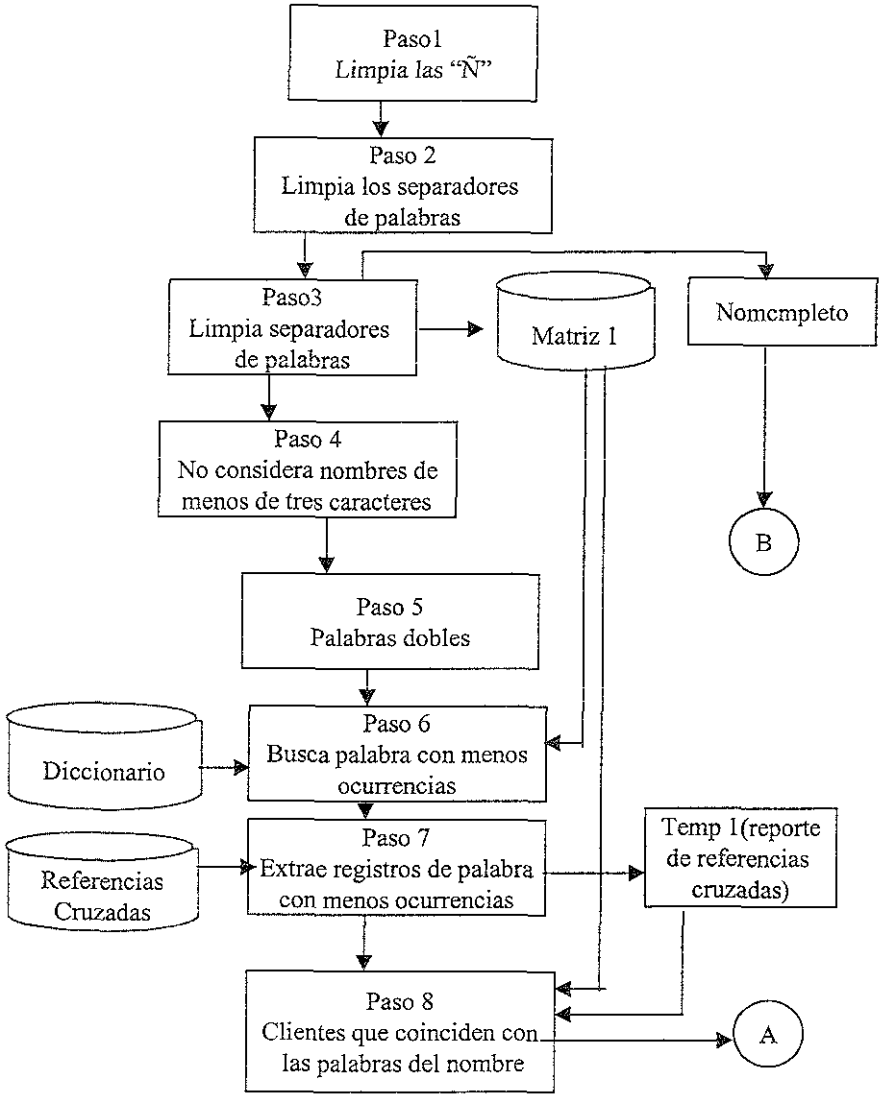

Figura 3.3 Algoritmo para la búsqueda de homónimos.

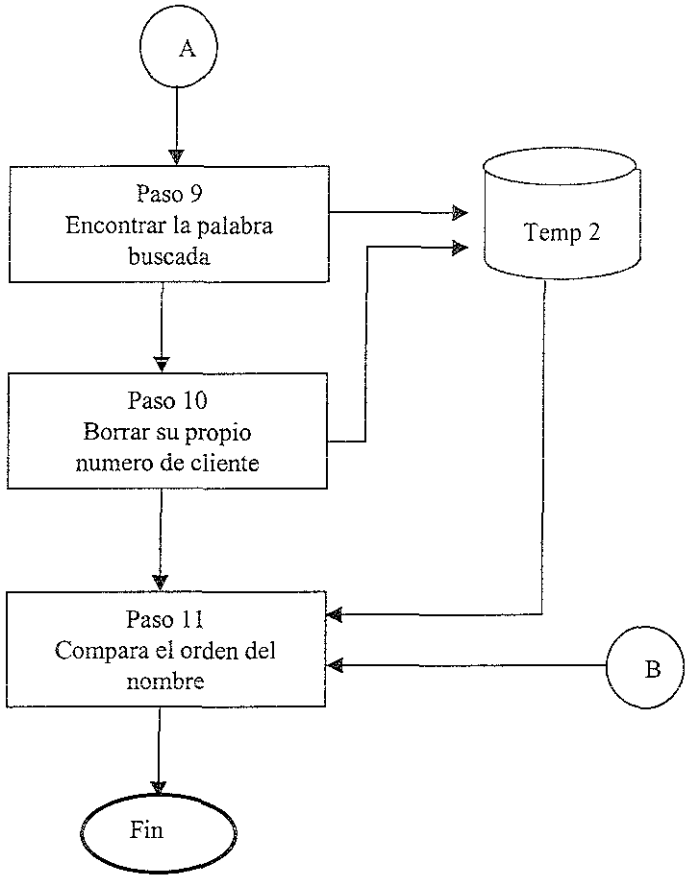

Figura 3.3 continuación

Tabla 3.7 Resultado de la búsqueda del nombre ANGELES ACUÑA MARTINEZ

ID CONSECUTIVO	ORDEN	NUMERO DE CLIENTE
4254 (ACUÑA)	1	785643
4254 (ACUÑA)	3	354466
9944 (ANGELES)	4	354466
9944 (ANGELES)	4	898871
9944 (ANGELES)	5	4493548
9944 (ANGELES)	6	8873422
9944 (ANGELES)	6	785643
80061 (MARTTINEZ)	2	785643
80061 (MARTTINEZ)	6	354466
80061 (MARTTINEZ)	1	4493548

Para mayor facilidad ordenamos la tabla por la tercera columna – numero del cliente.

Tabla 3.8 Resultado de la búsqueda ordenada por número de cliente.

ID CONSECUTIVO	ORDEN	NUMERO DE CLIENTE
80061 (MARTTINEZ)	6	354466
4254 (ACUÑA)	3	354466
9944 (ANGELES)	4	354466
9944 (ANGELES)	6	785643
80061 (MARTTINEZ)	2	785643
4254 (ACUÑA)	1	785643
9944 (ANGELES)	4	898871
80061 (MARTTINEZ)	1	4493548
9944 (ANGELES)	5	4493548
9944 (ANGELES)	6	8873422

Ejemplo: Si buscamos un nombre que tiene tres palabras diferentes, él número de este cliente debe de aparecer también tres veces dentro de la tabla de trabajo. (Ver tabla 3.8)

Debido que el nombre buscado tiene tres palabras nos interesan los números del cliente que aparecen tres veces. Estos son 354466 y 785643. Puede darse el caso que si tenemos apellidos repetidos, entonces el algoritmo va a buscar las palabras que aparecen tres o más veces. Esto es porque, vamos a encontrar los registros dos veces por cada palabra. El algoritmo va a reportar que el nombre buscado se parece a estos dos nombres encontrados. Para simplicidad ignoramos las palabras de menos de tres caracteres como son; "DE", "LA", "MA" y "MTZ".

Paso 10:
En este paso se borra de la tabla el resultado TEMP2, él numero del cliente nuevo, si es que existe. Esto se hace porque, no tiene caso reportar como encontrado a su propio numero de cliente buscado.

Paso 11:
Falta por determinar el orden en que los apellidos aparecen, dado que no es lo mismo "Medina Pérez José" que "Pérez Medina José". Para esto usamos la variable NOMCOMPLETO del paso tres en la que se guarda el orden en el que aparecen los apellidos. Ahora por cada cliente candidato a ser igual, armamos una cadena del apellido paterno, materno y el nombre separados por un blanco. Recordamos que la regla es que los apellidos aparezcan siempre en el orden Apellido paterno, Apellido materno y nombre. Esta cadena la estamos buscando si aparece como una subcadena de la variable temporal NOMCOMPLETO donde guardamos el nombre. Si no encontramos la subcadena dentro de la variable, entonces esto significa que el apellido paterno y el materno están al revés, por lo tanto no lo consideramos como una posible solución y eliminamos este cliente de la tabla resultante. El resto de clientes los guardamos como posibles parejas al nuestro cliente que buscamos. En todas estas parejas los nombres coinciden, pero no sabemos si es la misma persona o se trata de homónimos.

3.4.3. Descripción del algoritmo para determinar personas iguales

En la siguiente fase de la integración tenemos que determinar por cada pareja de clientes que se llamen igual, si son la misma persona o no. Para esto tenemos que recurrir a la información detallada, que se encuentra dentro de la base de datos. Tenemos que recurrir al RFC o al domicilio del cliente utilizando la parte del código postal en caso de que no tengamos la información sobre el RFC. Si los RFCs y los nombres de dos personas coinciden, esto se considera mas que suficiente para comprobar que se trata de la misma persona. De manera más estricta deberíamos de comparar la dirección, en la parte de nombre de la calle, numero de departamento, etc. Debido a que las direcciones no están bien estructuradas el algoritmo de comparación será muy complicado. Utilizando para la comparación únicamente el código postal nos da un margen de error más grande, pero para el propósito del algoritmo consideramos que esto es suficiente, ya que la toma de decisión no va a ser automática si no que estará sujeta a revisión manual. Además cuando encontramos que se trata de las mismas personas tenemos que decidir cual de los dos clientes se va a integrar al otro o más bien quien se queda en la base de datos y quien desaparece. El cliente que se quede tendrá que absorber todos los atributos del cliente que desaparece.

La entrada del algoritmo son las parejas de números de clientes homónimos obtenidos en la fase anterior. La salida del algoritmo esta conformada de dos archivos. El primer archivo consta de las parejas de clientes que son la misma persona y deben de ser integradas sin revisión (la decisión es automática). El segundo archivo contiene información detallada de aquellas parejas que por falta de datos o por datos erróneos no pudieron ser determinadas como la misma persona con una certeza absoluta de 100%. Este archivo se debe de analizar visualmente por el personal capacitado para la toma de decisión. (Ver figura 3.4)

Paso 1:
El algoritmo empieza con leer una pareja de clientes y limpiar todas las variables de trabajo. Por cada cliente de la pareja se extrae de la base de datos de clientes la información necesaria para el funcionamiento del algoritmo. La información esta compuesta por el RFC,

tipo de persona (física o moral), el código postal de la dirección y los productos o servicios que tiene este cliente con la empresa.

Paso 2:
En la comparación tenemos que distinguir el caso cuando los dos RFC son disponibles y cuando uno de los dos RFC falta. Cuando los dos RFC son disponibles y son iguales se trata de la misma persona. Para comparar los RFC de personas físicas se comparan las primeras diez posiciones del mismo. Las primeras cuatro posiciones, que son alfabéticas y las siguientes seis posiciones, que son numéricas y representan la fecha de nacimiento, en formato año/mes/día. En el caso de personas morales se comparan nada más las primeras nueve posiciones. Las primeras tres son alfabéticas y las siguientes seis representan la fecha del acta constitutiva de la empresa en formato año/mes/día. El algoritmo continua con el paso 6. Cuando los dos RFC son disponibles y no son iguales, entonces la pareja se rechaza y el algoritmo se regresa al paso 1..
En el caso de que solo uno de los RFC esté disponible, entonces continuamos con el paso 3.

Paso 3:
En caso de que falte uno de los dos RFC extraemos de la BD los productos asociados con cada uno de los dos clientes. Comparamos si entre los productos se encuentra el producto "Tarjeta de débito" (código 0267). Si este producto existe el algoritmo pasa al paso 4.
Si no se encuentra este producto recurrimos a la dirección y comparamos los códigos postales. Si alguno de los códigos postales no tiene valor correcto o con otras palabras, viene sin valor (nullo), con valor en blanco, o contiene "00000", rechazamos la pareja y continuamos con el paso 1.
Si los dos códigos postales tienen valor real y son iguales, grabamos la pareja de clientes dentro del archivo para revisión manual y continuamos con el paso 1. Es muy probable que se trate de la misma persona puesto que vive en la misma zona, pero por la falta de RFC, necesitamos sujetarlo a revisión manual. Si los códigos postales no son iguales rechazamos la pareja y regresamos a paso 1.

Paso 4:
En el caso de que uno de los clientes no tenga ni RFC, ni dirección, revisamos los productos que los clientes tienen contratados con la empresa. La empresa tiene dos productos que se llaman relacionados. Uno es la tarjeta de débito y el otro es la cuenta de cheques relacionada con esta tarjeta de débito. Se efectúa la búsqueda en la base de datos de clientes sobre la relación entre las dos cuentas. Si determinamos que el producto de un cliente de la pareja es tarjeta de débito y el producto del otro cliente es una cuenta de cheques y además si determinamos que la cuanta de cheques y la tarjeta de débito están relacionadas, podemos estar casi seguros que se trata de la misma persona y seguimos con el paso cinco. En caso contrario se rechaza la pareja y se sigue con el paso 1.

Paso 5:
Todavía queda una duda por aclarar. La tarjeta de débito tiene un numero limitado de caracteres para grabar el nombre, por esta razón en la mayoría de los casos los nombres en la tarjeta de débito se abrevian. En este momento entramos en un algoritmo que nos permite encontrar si los dos nombres son iguales al 100% o tiene algunas palabras iguales y el resto

de palabras tienen nada mas iguales la primera letra por ser abreviada. Aquí vamos a poner la condición de que debemos de tener iguales como mínimo dos palabras completas, no abreviadas. El algoritmo es sencillo, pues coloca las palabras de cada nombre en una tabla de trabajo separada. Las tablas se llaman "tab1" y "tab2". Después comparamos las dos tablas y eliminamos las palabras iguales de cada tabla. Quedan únicamente palabras enteras o abreviadas. De esta manera comparamos si las palabras abreviadas son iguales a la primera letra de las palabras enteras. Aquí recordamos que nuestros algoritmos iniciales no funcionan con palabras de una sola letra, por esto tenemos que validar la información.

Paso 6:
Cuando ya estamos seguros que los dos clientes son la misma persona necesitamos determinar cual cliente se queda y cual desaparece. Para tomar esta decisión nos basamos en los productos que cada cliente tiene en la empresa. Tenemos una tabla que contiene la jerarquía de los productos. El cliente que tiene un producto de mayor jerarquía se tiene que quedar. Si los dos tienen la misma jerarquía se toma la decisión, para que se quede el cliente con mas antigüedad y su número de cliente sea el menor de los dos. El algoritmo continua con el paso 1.

Paso 7:
Este paso se efectúa cuando en el paso 1 se encuentra la condición de "fin de archivo". Ya tenemos determinadas las parejas de padres e hijos para efectuar la integración. Durante la integración el hijo desaparece y sus cuentas se integran al padre. Esto se hace con la finalidad de integrar todos los clientes duplicados y dejar uno solo. Durante el proceso se puede presentar una situación donde el mismo cliente, puede ser determinado como hijo en una pareja y también como padre en otra pareja. Si se procesa primero la relación donde el cliente apárese como hijo, este se integrara y desaparecerá como cliente, por lo tanto la siguiente integración donde el mismo cliente aparece como padre no se podrá efectuar debido a que el cliente ya no existe (fue integrado anteriormente). Por esto se deberían de buscar todos aquellos clientes hijos que se encuentran también como padres en otras relaciones. Utilizamos la propiedad transitiva: si A=B y B=C, entonces A=C. Separamos en una tabla de trabajo todos los clientes que aparecen de los dos lados de la relación. En dirección de arriba hacia abajo empezamos a hacer las sustituciones necesarias.

Ejemplo de clientes que aparecen en los dos lados de la relación en diferentes parejas.

A	**B**
D	E
G	X
B	**C**
R	Z

Se obtiene las relaciones, donde el cliente B aparece en los dos lados de la relación en diferentes parejas.

A	**B**
B	C

Se sustituye el cliente que aparece del lado izquierdo de la relación por su padre de la otra relación donde el mismo cliente aparece del lado derecho.

A	B
A	C

Con esto concluimos el algoritmo de determinar parejas de clientes por integrar.

Las ventajas de los algoritmos presentados son varias. En primer lugar la realización del algoritmo es muy económica. Además de que tiene un precio de desarrollo y ejecución diaria bajo, el beneficio es casi inmediato, lo que significa que el resultado del algoritmo es efectivo casi de inmediato. El tiempo desde la captura de un nuevo cliente hasta su integración es menos de veinticuatro horas. Otra ventaja es el concepto del diccionario que se podría utilizar, para propósitos generales. El diccionario refleja de manera inmediata todos los datos de mala calidad introducidos en la base de datos. Puede constituir un núcleo y alrededor de él se pueden construir otros sistemas para la limpieza de datos. Otra ventaja es que el algoritmo trabaja con palabras separadas y no con los nombres completos. Esto significa que le orden de las palabras no tiene importancia desde el punto de vista de la calidad.

3.4.4. Sugerencias para la implementación de los algoritmos en la base de datos

Los algoritmos están orientados para trabajar con una base de datos relacional con lenguaje SQL. Con pequeñas modificaciones se puede adaptar para cualquier sistema de clientes basada en otro ambiente, como archivos VSAM o base de datos no relacionales. Las siguientes sugerencias están orientadas, para la implementación en un sistema abierto UNIX y un manejador de base de datos Sybase.

Requerimientos de hardware.

Se sugiere que el sistema corra en equipos UNIX. Por ejemplo un equipo HP-9000 con el sistema operativo HP-UX versión 10.20 y posteriores Se necesitan los programas del sistema operativo awk, join y sort. Estos comandos se pueden usar para el algoritmo de creación de la base de datos. El comando sort necesita áreas de trabajo en disco. El tamaño del área depende de la cantidad de registros que contiene la base de datos de clientes. Se necesitan alrededor de doscientos megabytes de espacio libre en disco.

Requerimientos de software.

Se sugiere que el sistema corra con la base de datos de Sybase Adaptive Server Enterprise versión 11.5. Se podrían usar los programas de Sybase isql y bcp. El programa bcp se usa para extraer los nombres de la base de datos de clientes y también se usa para la carga de los datos en el diccionario. El programa isql se usa para correr los comandos en le servidor de la base de datos. Si los algoritmos programados van a correr en el mismo ambiente donde se encuentra la base de datos de clientes es recomendable que el diccionario y los procesos tengan una propia base de datos. Esto es para no impactar el desempeño de la base da datos de clientes mientras están corriendo los procesos de integración. El tamaño de esta

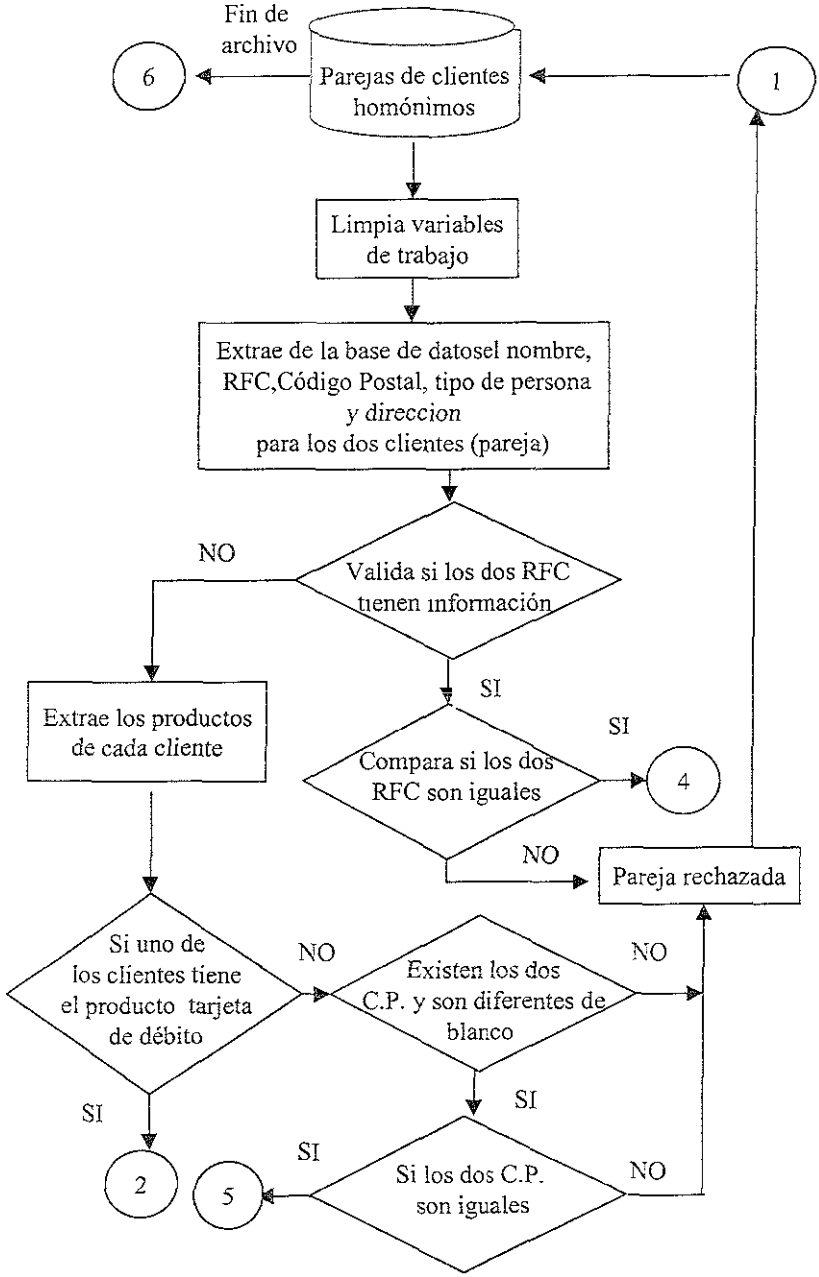

Figura 3.4. Algoritmo para determinar personas iguales.

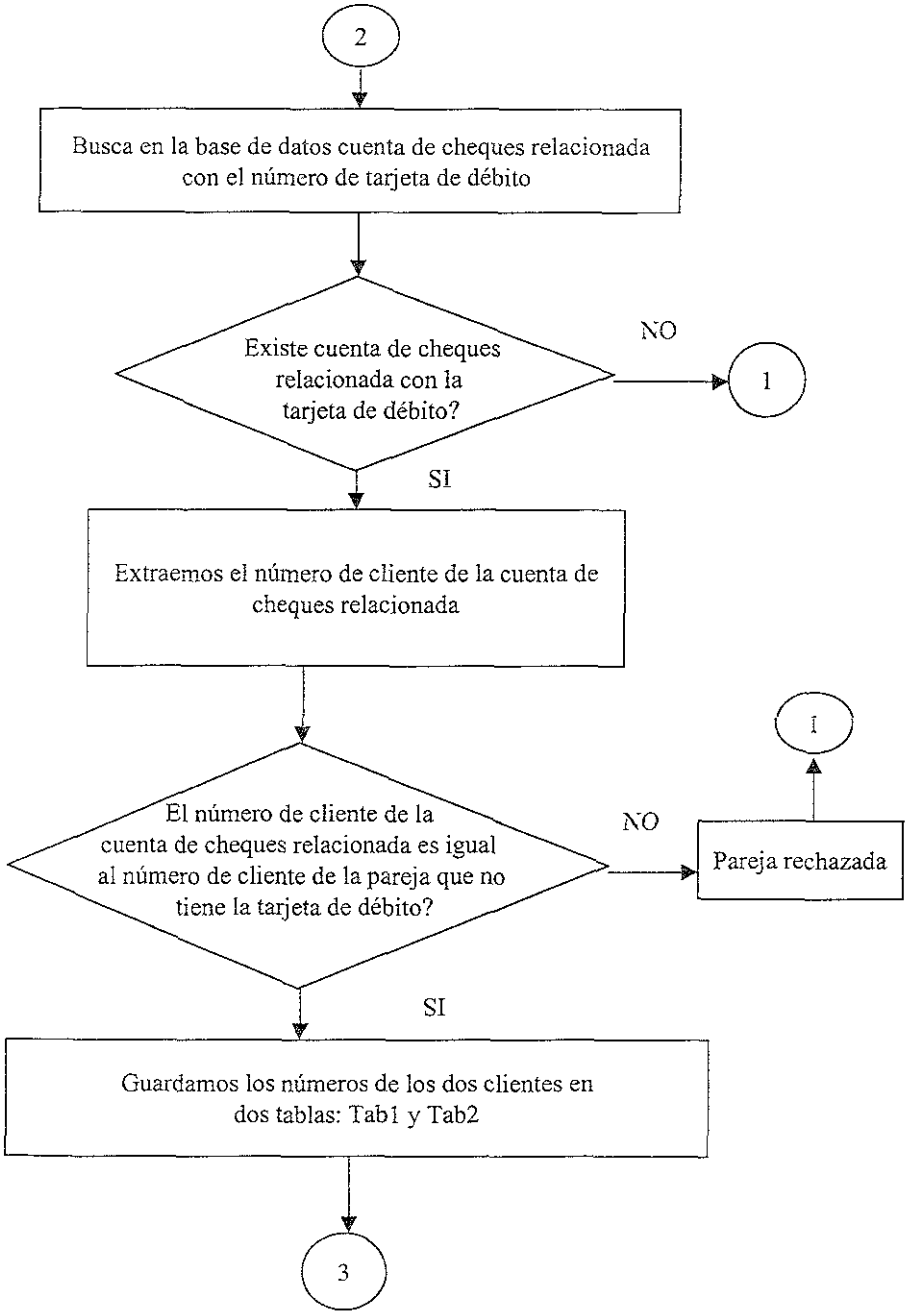

Figura 3.4. continuación

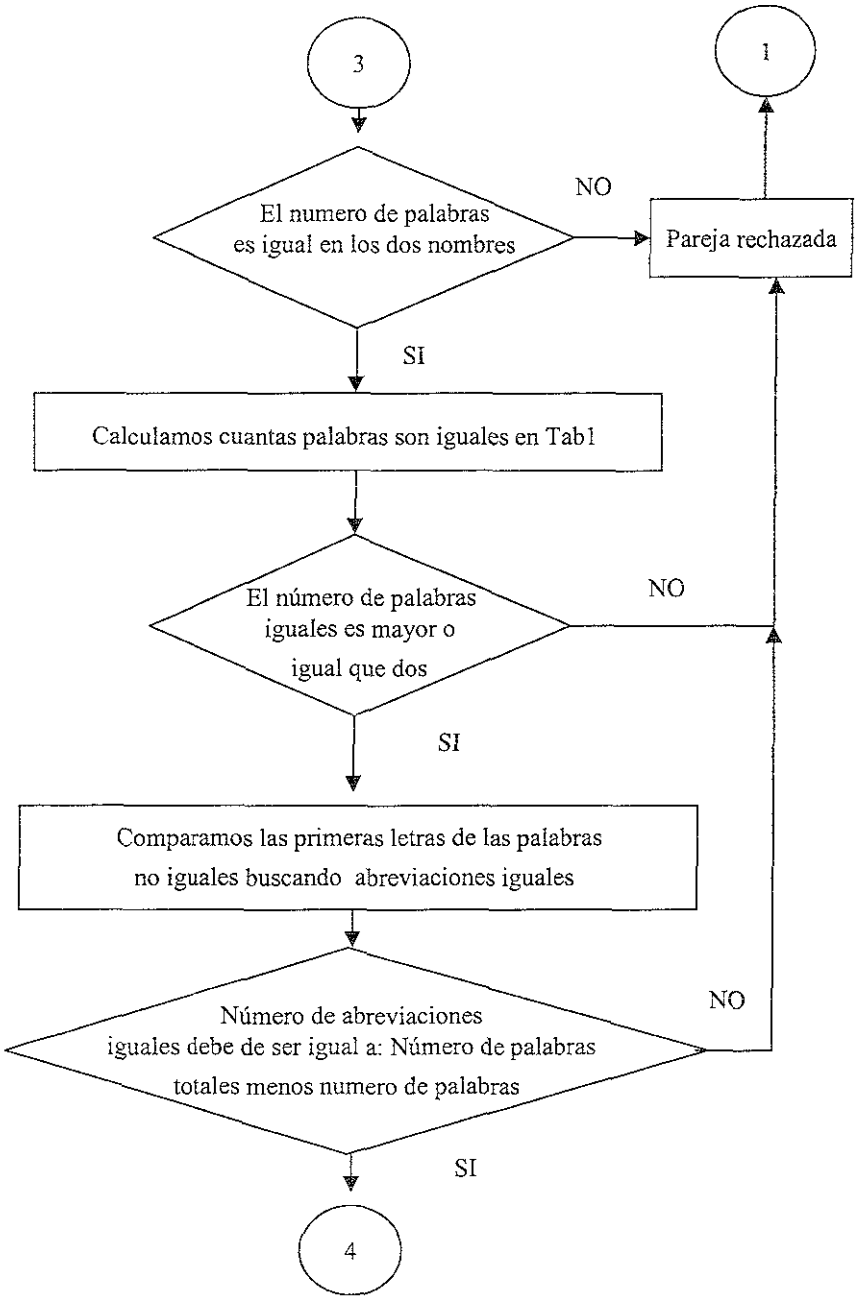

Figura 3.4. continuación

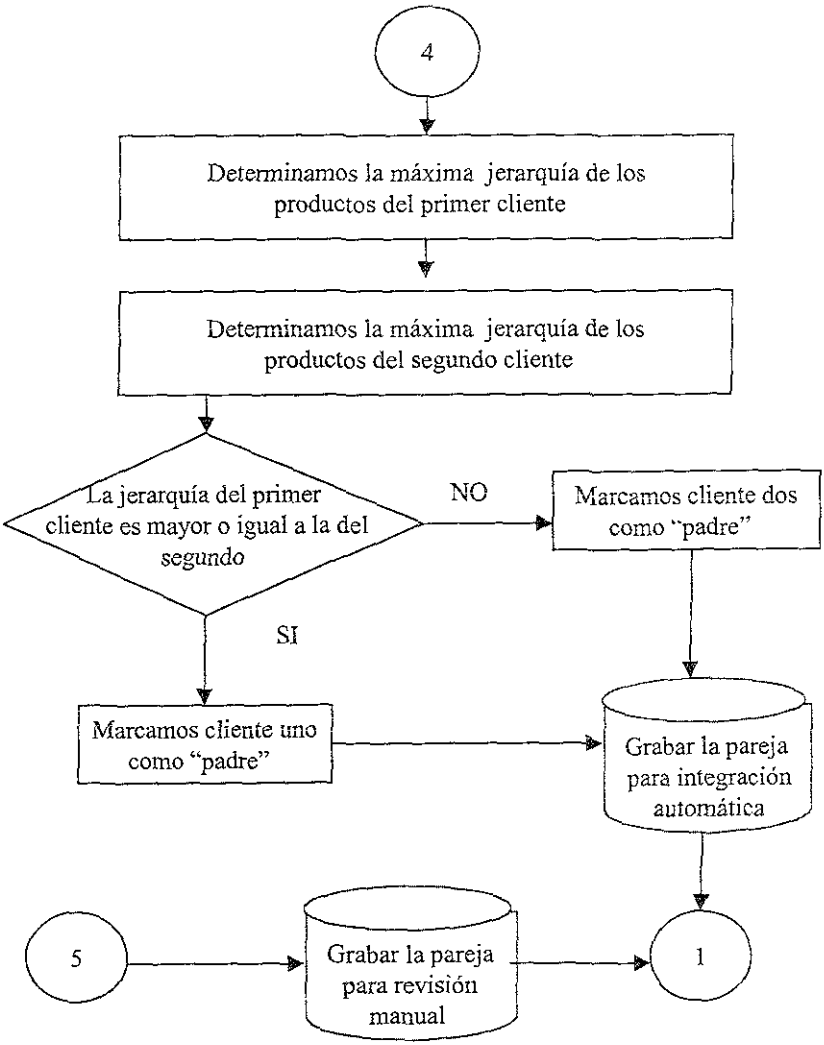

Figura 3.4. continuación

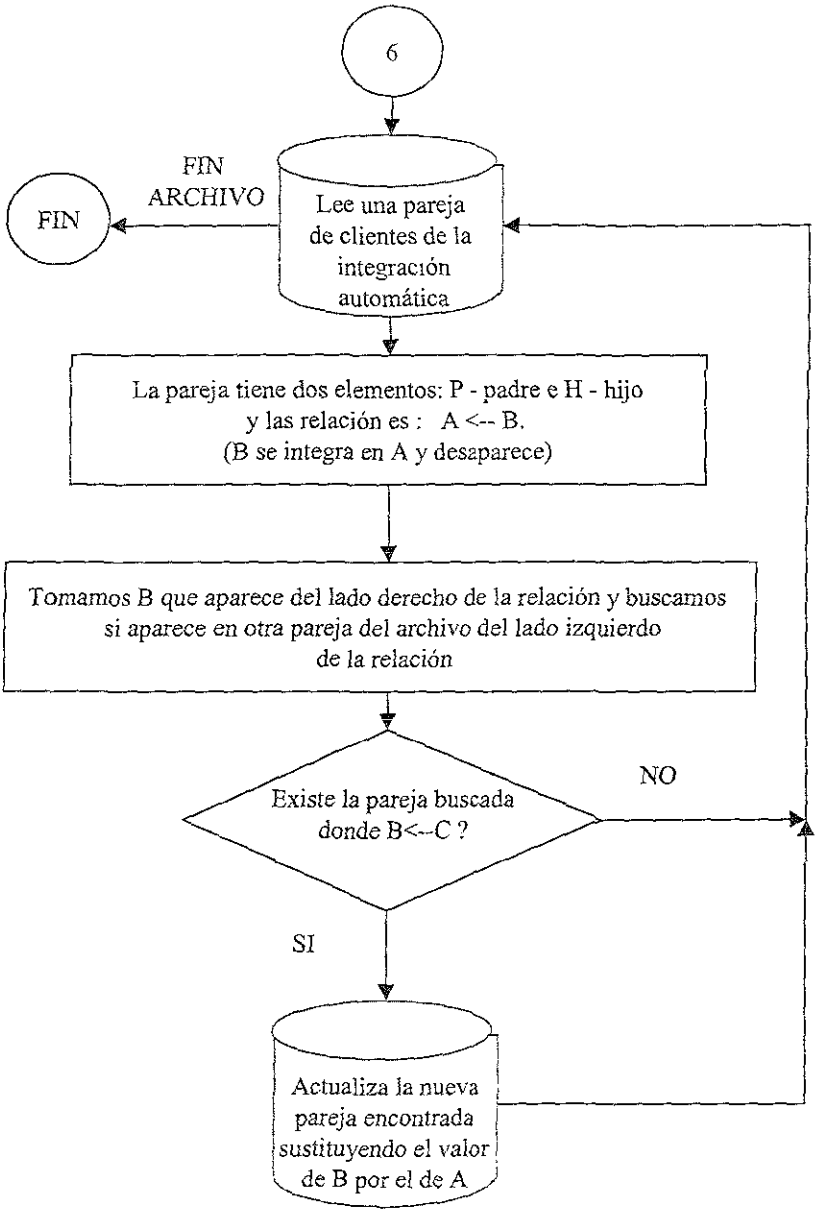

Figura 3.4. continuación

base de datos también depende del volumen de la información. Por ejemplo para una base de datos de tres millones de clientes el tamaño del diccionario será aproximadamente de ciento treinta mil palabras mismas que requieren de solo nueve megabytes de espacio. Las referencias cruzadas por su lado requieren de trescientos noventa megabytes de espacio. El proceso de integración requiere áreas temporales de trabajo (la base de datos "tempdb") de un mínimo de cien megabytes de memoria.

Secuencias sugeridas para la instalación del sistema

1. El primer paso es crear la base de datos para el diccionario con un tamaño de quinientos megabytes de datos más setenta megabytes para el log de transacciones.
2. Validar el tamaño configurado de la base de datos de trabajo del servidor Sybase "tempdb" e incrementarlo según sea necesario, para asegurar un mínimo de cien megabytes para los procesos. Los cien megabytes permitirán la corrida simultánea de un promedio de seis procesos. Si se requiere la corrida de procesos simultáneos adicionales, se debe de prever el incremento del tamaño de la "tempdb".
3. Se crean las tablas del diccionario que se llaman "diccionario" e "diccionario_índex". El tamaño no tiene importancia porque se adapta a los requerimientos de la información.
4. En la base de datos del diccionario se carga el procedimiento almacenado (stored procedure), que ejecuta el algoritmo de homónimos.
5. En UNIX se cargan los dos "scripts"- secuencias de comandos. El primero es para la creación del diccionario, el segundo para la ejecución del proceso de integración.

Recomendación para la creación del diccionario.

Dependiendo de los cambios que sufre la base de datos de clientes, la creación del diccionario se puede correr diario o semanal. Se debe de considerar que según el volumen de los datos la creación puede durar de una hora hasta una hora y media. Durante la creación del diccionario es primer paso es la extracción de tres archivos respectivamente para apellido paterno, materno y nombre. Para cada parte del nombre se lanza un proceso. Estos proceso pueden ser lanzados uno tras otro o en paralelo. Es recomendable que se haga en paralelo porque el tiempo de extracción sea tres veces menor.

Recomendación para la corrida del proceso de integración.

El sistema permite la corrida simultanea de varios procesos. Para recortar el tiempo total del proceso de integración se sugiere que este sea dividido en varios procesos que corran en paralelo.

CONCLUSIONES

La calidad de los datos y de la información es un indicador de cómo los recursos informáticos soportan la necesidad de información de cualquier empresa.
Los recursos de datos en la mayoría de las empresas se caracterizan por una calidad dispareja de la información. Calidad que no responde a las demandas del negocio. La principal razón de esto son los sistemas hereditarios que existen dentro de la empresa. Estos sistemas crean y almacenan los datos para cumplir con funciones específicas, pero en ediciones limitadas, con poca validación y sin ser debidamente documentados. Con el proceso natural de cambio y reemplazo de los viejos sistemas, los conocimientos de la calidad que se tienen pueden desaparecer si no están documentados.
Es un hecho que mientras los recursos informáticos aumentan de tamaño y complejidad, la calidad de los datos y la información disminuye. Cada vez mayor cantidad de personas desarrollan más y más sistemas, capturando mayor cantidad de datos con diferentes técnicas y con menos control sobre los mismos. El desarrollo de los sistemas de información y las aplicaciones de arquitectura cliente - servidor sin una arquitectura en común, junto con la falta de un formal criterio sobre la calidad de los datos y la información hacen la situación todavía peor. Las empresas necesitan tomar el control de sus datos comprendiendo el concepto y los criterios de calidad, ligados con la demanda del negocio. Hacer esto es una tarea complicada y difícil de resolver, pero no es imposible si se hace dentro de una arquitectura conjunta de datos.
La calidad de datos consiste en que estos deben de ser integrados, precisos y completos. La integridad de los datos trata sobre la forma en que los datos son mantenidos dentro de los recursos informáticos. Esto consiste en tener un criterio para definir una estructura apropiada y mantener sus valores correctos, permitiendo así una retención de los datos cruciales para tener a tiempo los datos derivados. De la precisión de los datos depende la manera en que estos reflejaran el mundo real. Que los datos sean completos se le llama al proceso que implica asegurarse que todos los datos necesarios cumplan con la demanda de información y existan dentro de los recursos informáticos. Uno de los aspectos más importantes de la calidad de los datos y la información es mejorar los datos actuales por medio de análisis que ajustan los datos para asegurarnos que representen el mundo real.
Otro aspecto importante es determinar su linaje y su transformación por medio de procesos sucesivos. Los datos pueden ser identificados como primarios o secundarios o como recursos oficiales y no oficiales para determinar de donde se originan y donde está localizado el registro de referencias. Manteniendo el componente del tiempo en todas las versiones de los datos y bases de datos temporales es otro aspecto de la calidad a medida que cada vez mayores cantidades de información son almacenadas para su análisis de tendencias y proyecciones. También es de suma importancia actualizar los múltiples recursos de datos para mantenerlos sincronizados por medio de técnicas retroactivas. La calidad de los datos y la información refleja el nivel de la empresa y mantenerla es siempre el sinónimo de tener control sobre el negocio. La calidad facilita la captura de los datos recientes que ayuda la retroalimentación para apoyo de los datos ya existentes. Para mejorar la calidad es necesario, entender el nivel actual, definir el nivel necesario de calidad al que se quiere llegar y cambiar los datos siempre teniendo el control sobre los procesos necesarios para lograr un resultado óptimo.

Los problemas de calidad de datos ocurren cuando la información no cumple con las expectativas del usuario o cliente. La calidad de la información ya no es una noción abstracta. El hecho de que la calidad se puede medir y se le puede dar un valor la convierte en algo tangible y más fácil de entender. Los productos para la calidad de la información, son herramientas poderosas en la lucha contra los problemas de la calidad de la información. El éxito de uso de estos productos requiere entender los problemas de negocio que se quieren resolver, así como su funcionalidad y sus limitaciones. Entender como implementar estos productos dentro de los procesos de calidad de la información.

Las técnicas para la mejora de la calidad son actividades que apoyan la empresa en analizar, medir, documentar, limpiar, prevenir y controlar la calidad de la información. Medir la calidad de la información no es para crear un reporte si no para incrementar la satisfacción del cliente y la eficacia del negocio eliminando problemas causados por la mala calidad de la información. La implementación de una herramienta para la calidad requiere la formación de un equipo de personas altamente capacitadas de diferentes áreas de la empresa: áreas de negocios, áreas de atención al cliente, y áreas de ingeniería de sistemas. Ahora después de tener definido el proceso, para medir y mejorar la calidad de la información y después de adquirir herramientas y técnicas que soportan estos procesos y después de armar el equipo de trabajo, se puede asegurar que la empresa va a efectuar un cambio cultural, y hacer que la calidad suceda y se convierta en una herramienta del negocio.

BIBLIOGRAFÍA

Ralph Kimball, The Data Warehouse Toolkit: Practical techniques for building data warehouses.
New York: John Wiley, 1997

Michael Bracket, Data sharing a common data arquitecture
New York: John Wiley, 1997

Len Silverston, W. H. Inmon, Kent Graziano, The data model resourse book
New York: John Wiley, 1998

Buckland J. R. Fowinkle, L. Shroyer: Total Quality Management in Information Services
New York: Wiley Interscience, 1997

English L. P. Data quality: Meating Customer Needs, DM Review, November 1996

English L. P. Data Stewardship: A Human Solution of Data Integrity, Database Programing & Design. April 1993

Huang K-T, Yang Lee & Richard Wang. Quality Information and Knowledge
Upper Saddle River: Prentice Hall, 1999

Redman T. Data Quality for the Information Age.
Boston: Artech House, 1998

Reingruber, M. C. and William W. Gregory. The Data Modeling Handbook: A Best-Practice Approach to Building Quality Data Models.
New York: John Willey & Sons, 1994

Walton, M. Tha Daming Management Method.
New York: Putnam Publishing Group, 1998

Wang, R. Lee Yang, Leo Pipino, and Diane Strong: Mange Your Information as a Product.
MIT Sloan Management Review, 1998

http:// www.datawarehouse.com

http.// www.infoimpact.com

http:// www.datawarehousing.com

http:// www.infomanager.fi