Contenido

1 UTILIDAD DEL DIAGNÓSTICO DE LA CALIDAD DE LOS DATOS 1

2 DIAGNÓSTICOS PARA ESTIMAR LA CALIDAD DE LOS DATOS 3

2.1 ASPECTOS CONTEXTUALES 3
2.2 DISEÑO DE LA BASE DE DATOS 5
2.3 INSPECCIÓN DE INTERFACES CON EL USUARIO 7
2.4 PRUEBAS DE DIAGNÓSTICO 7
2.5 CRITERIO DE USUARIOS 9
2.6 CALIFICACIÓN DEL DIAGNÓSTICO DE DATOS 9
2.7 EVALUACIÓN DEL DIAGNÓSTICO 12

3 ASPECTOS POR CONSIDERAR 13

3.1 CONSIDERACIONES AL GESTIONAR UNA ADQUISICIÓN DE SISTEMAS 13
3.2 CONSIDERACIONES AL DESARROLLAR SISTEMAS 13
3.3 CONSIDERACIONES AL MIGRAR DATOS 15

4 TESTIMONIO DE EXPERIENCIAS 17

4.1 ELABORACIÓN DE UN MODELO DE DATOS CORPORATIVO 17
4.2 DESARROLLO DE UN SISTEMA 17
4.3 UN PROYECTO DEL Y2K 18

5 CONCLUSIONES Y RECOMENDACIONES 19

5.1 RECOMENDACIONES 19

ANEXO 20

CORRUPCIÓN DE LA INFORMACIÓN: EL ASESINO SILENCIOSO DEL ERROR DEL AÑO 2000 20

BIBLIOGRAFÍA 22

Tablas

TABLA 1 : EVALUACIÓN DEL DIAGNÓSTICO DE CALIDAD DE DATOS 9
TABLA 2: EVALUACIÓN DEL DISEÑO DE BASES DE DATOS 10
TABLA 3: MÉTRICA PARA EVALUAR INTEGRIDAD REFERENCIAL 11
TABLA 4: MÉTRICA PARA EVALUAR SEÑAS DE DIRECCIÓN DE COBRO 11

1 Utilidad del diagnóstico de la calidad de los datos

El diagnóstico de la calidad de datos es una herramienta para mejorar los resultados en la ejecución de las actividades que se describen a continuación.

Diagnóstico de los sistemas

Según nuestra experiencia, una señal del estado de corrección de un sistema puede ser deducida de la calidad de los datos, pues éste es un valor directamente proporcional a la calidad del código de programación.

Implantación de sistemas

Implantación es la fase que pone en operación un sistema una vez que su programación ha sido terminada y probada. Si el sistema inicia su operación con datos nuevos, un diagnóstico asegura que la calidad de los datos que el sistema genera cumple con las expectativas antes de ponerlo en operación. Si el sistema inicia su operación a partir de datos existentes, el diagnóstico es útil para:

- Especificar la carga de datos.
- Dimensionar el volumen de los datos que se deben reparar.
- Dimensionar el volumen de los datos desconocidos.
- Dimensionar el volumen de los datos que se deben capturar desde medios no electrónicos.
- Estimar los recursos para la migración de datos.

Escalamiento

Escalar una aplicación es aumentar su alcance a un contexto más grande, por ejemplo, cuando la funcionalidad se extiende de una empresa a toda la corporación. Si existen problemas de calidad de datos, un escalamiento de la aplicación los magnificará.

Un diagnóstico en la instalación inicial ayuda a dimensionar el riesgo del escalamiento. Un diagnóstico en las áreas donde será extendido el alcance ayuda a dimensionar los recursos requeridos para hacer el escalamiento.

Mantenimiento

En cuanto a los recursos requeridos para realizar mantenimiento, el diagnóstico ayuda a estimar:

- Tiempo de detección de fallas, cuando éstas son causadas por errores de datos.
- Recursos para la reparación de fallas, basado en la calidad del código que es inferida de la calidad de los datos.

Integración de sistemas

Al integrar sistemas[1], es necesario diagnosticar la calidad de los datos que se comparten con el fin de:

- Especificar filtros para asegurar compatibilidad de formatos y códigos.
- Especificar criterios de aceptación de datos.

Adquisición

La calidad de los datos debe ser un factor de evaluación en los términos de referencia para la adquisición de una aplicación.

Diagnóstico de la calidad de los procesos de la empresa

El diagnóstico de la calidad de datos, no necesariamente almacenados en medios electrónicos, provee señales acerca de la cultura de calidad de la empresa como un todo. La calidad de los datos es el resultado de procesos empresariales que se conciben y realizan con calidad.

[1] Por ejemplo: facturación con inventario, o planillas con contabilidad.

2 Diagnósticos para estimar la calidad de los datos

Esta sección propone un método para la realización y evaluación de un diagnóstico de datos. La evaluación se realiza desde las siguientes pespectivas:

Conceptos de evaluación	Justificación
Aspectos contextuales	La presencia de estos indicios revela riesgos de calidad en los datos
Diseño de la base de datos	Un buen diseño reduce el riesgo en la calidad de los datos
Interfaz con el usuario	Una interfaz adecuada, no ambigua, que valide los datos de entrada consistentemente y sintonizada con el contexto del usuario tiende a incrementar la calidad de los datos
Realización de pruebas de diagnóstico	Proveen evidencias documentales acerca de la calidad de los datos
Criterio de usuarios	Su criterio completa la visión de los anteriores conceptos.

Hay dos conceptos de evaluación cuya medición puede ser subjetiva, estos son: Aspectos Contextuales y Criterio de Usuario. Un diagnóstico riguroso podría obviar estos conceptos en la medición de la calidad, aunque siempre pueden ser considerados como elementos que provean alertas que ayuden a perfilar las pruebas de diagnóstico y determinar qué inspeccionar en la base de datos. En este sentido, puede ser útil realizar un diagnóstico general antes de uno detallado. Un diagnóstico general incluiría: Aspectos contextuales, Criterio de usuarios y una inspección *no exhaustiva* del diseño de la base de datos. El diagnóstico detallado estará dirigido hacia confirmar y aislar lo encontrado en el diagnóstico general.

Más adelante se propone un instrumento para realizar la evaluación integral basado en una ponderación de la evaluación de cada concepto.

2.1 Aspectos contextuales

Los aspectos contextuales son aquellos indicios que revelan que una aplicación tiene problemas de calidad de datos fácilmente observables.

Estos aspectos ayudan a perfilar las pruebas de diagnóstico, que se mencionan más adelante.

Existencia de un sistema paralelo

La existencia de un sistema paralelo que realiza funciones complementarias o correctivas del sistema actual, es un síntoma de alguna de las siguientes condiciones:

- la calidad de los datos del sistema actual es no satisfactoria
- el sistema actual no satisface todas las necesidades actuales
- el sistema actual tiene deficiencias de integridad

Integración de los sistemas

Sistemas no integrados tienden a decrementar la calidad debido a :

- Redigitación de datos, pues hay riesgo de que las reglas de validación sean diferentes en cada contexto.
- Redundancia en la representación de los datos, pues hay riesgo de generar inconsistencia.

Velocidad de crecimiento de la empresa

La velocidad de crecimiento de la empresa representa un riesgo pues, de manera típica, se presentan las siguientes situaciones:

- Muchas soluciones tecnológicas se improvisaron.
- La empresa no tiene la suficiente madurez para el establecimiento de estándares adecuados.

Tecnología de las plataformas

Una plataforma para la gestión de los datos basada en un Sistema Administrador de Bases de Datos (SABD)[2] es un indicador de menor riesgo que una plataforma basada en un sistema de archivos[3] o manejo de archivos planos[4].

En una plataforma basada en un SABD, se facilita hacer pruebas de calidad y de rendimiento.

Estandarización de códigos

Un sistema de información y la organización que lo rodea son espejos uno del otro. La organización refleja la calidad del sistema y el sistema refleja la calidad del manejo de información de la organización. Si la organización no ha estandarizado sus códigos, p.e.: código de género (femenino, masculino), códigos de sucursal, etc, o éstos no son de uso generalizado; entonces es posible que esta condición surja al inspeccionar los sistemas para su diagnóstico.

[2] Un SABD tiene los siguientes atributos: manejo de catálogo, reglas de seguridad, integridad, manejo del acceso concurrente, administración de transacciones y manejo de recuperación ante fallas.
[3] Por ejemplo: Fox Pro, Access, Clipper, dBase.
[4] Por ejemplo: manejo de archivos en Cobol, sin usar ninguna herramienta de administración y acceso a archivos.

Cantidad e impacto de fraudes en los últimos años

Esta es una señal muy clara de problemas en la calidad de datos pues refleja fallas en la detección oportuna de anomalías. Típicamente, un fraude requiere planeación y espera del momento oportuno, por lo que es una señal de que los problemas han existido por mucho tiempo.

Las cifras del sistema se pueden conciliar con fuentes de información emitidas desde otros sistemas

Para validar calidad de datos en una aplicación es importante verificar con una o más fuentes de conciliación independiente. Considérese una aplicación de Cobranza: la comprobación de sus cifras globales de procesamiento de entrada debe conciliarse con las emitidas por un Sistema de Tesorería; por otro lado, estas mismas cifras deben conciliarse con las existentes en el Sistema de Contabilidad. Una conciliación periódica de estas cifras es un buen indicador de calidad.

2.2 Diseño de la base de datos

Un buen diseño de la base de datos es un indicio positivo de la calidad de los datos. Los criterios de evaluación que se describen más adelante representan características de un buen diseño lógico para una base de datos típica. En un diagnóstico es importante encontrar señales de debilidad o fortaleza con respecto de estos criterios.

Normalización

El diseño es normalizado[5]. Si alguna sección del diseño es no normalizada, las razones para ello están documentadas.

Cantidad de atributos que aceptan nulos

Un riesgo con el uso de atributos que aceptan valores nulos es que crea la siguiente duda, ¿*el valor es nulo porque la información no aplica o porque no se conoce*?

Cantidad de atributos derivados[6]

Se incrementa el riesgo de que si cambian los atributos base no se recalcule el atributo derivado, con lo cual éste se torna inconsistente.

Integridad referencial

Todas las llaves foráneas están declaradas, las reglas de inserción y borrado en cascada son claras. La ausencia de integridad referencial es una clara señal de ausencia de calidad en el diseño físico de la base de datos.

[5] En el sentido de la teoría de normalización de bases de datos (usualmente relacionales).
[6] Atributos cuyo valor es el resultado de una expresión que referencia otros atributos en la base de datos.

Selección de llaves

La evaluación acerca de una selección adecuada de llaves toma en cuenta los siguientes aspectos:

- Unicidad del valor
- No nulidad
- El valor es fácil de obtener desde las fuentes de información
- Longitud: llaves excesivamente largas no son adecuadas.
- Si dos entidades están muy ligadas por una relación de dependencia, por ejemplo: Cuenta por cobrar y movimientos de una cuenta, entonces la llave de la primera debe facilitar el acceso a la segunda, por ejemplo: número de cuenta debe facilitar el acceso a la consulta de movimientos de tal cuenta.

Estructuración de información

Estructurar es representar la información contenida en un atributo en atributos más simples; entonces tenemos un atributo compuesto de otros atributos. La falta de estructuración puede dificultar el acceso y la actualización de datos.

Precisión de campos tipo *string*

La precisión es adecuada si el espacio para almacenar la información es suficiente para representar valores sin pérdida de calidad de datos. Por ejemplo: si el usuario debe recurrir a usar abreviaturas para escribir información en campos tipo *string*, esto es un indicador de que la precisión es insuficiente.

Precisión de campos numéricos

Aquí la precisión se mide en longitud del campo (¿es suficiente el tamaño para almacenar los valores promedio o pico?) y si la cantidad de decimales es suficiente para proveer corrección en los cálculos. Una señal de que la precisión tiene problemas es cuando al almacenar o acceder a los campos es necesario dividirlos o multiplicarlos por una potencia de 10.

Representación de códigos y enumeraciones

La representación de códigos está en la base de datos y no en el código del programa.

Reglas de integridad[7]

Las reglas de integridad centralizan la definición y ejecución de las validaciones sobre los atributos como una función del SABD: su uso ayuda a incrementar la calidad de los datos, al reducir problemas de consistencia.

[7] Este criterio es aplicable en sistemas de información corriendo sobre un administrador de bases de datos.

Tipos definidos por el usuario[8]

Estos tipos facilitan la redefinición de los valores que son aceptados como válidos y su formato de captura. Su existencia es una señal positiva acerca de la consistencia de los datos.

2.3 Inspección de interfaces con el usuario

Al inspeccionar interfaces (pantallas, informes) se puede encontrar evidencia de problemas de calidad en los siguientes aspectos:

- Precisión
- Formatos de captura inconsistentes para campos del mismo tipo
- Reglas de validación inconsistentes para campos del mismo tipo
- Atributos siempre nulos; esto es, atributos presentes en interfaces que nunca son capturados
- Atributos cuyo nombre no es consistente con la información que muestran

La inspección de interfaces permite reconocer si la interacción de la interfaz se acopla a las funciones realizadas por el usuario.

2.4 Pruebas de diagnóstico

El diseño de las pruebas de diagnóstico se dirige hacia la corroboración de las señales detectadas al inspeccionar los Aspectos Contextuales mencionados anteriormente.

Para cada prueba debe establecerse una definición de cumplimiento de calidad, por ejemplo: "el 99% de los datos sobre clientes tienen información completa de su dirección de cobro". Si el ambiente lo permite, la prueba debe realizarse sobre toda la población, esto es posible cuando la plataforma es un Sistema Administrador de Bases de Datos (SABD) que permite consultas expresadas mediante SQL; en otro caso, se recomienda realizar la prueba sobre una muestra.

Los pasos de las pruebas de diagnóstico se describen a continuación.

Determinar tamaño de la muestra

Para determinar el tamaño de una muestra se sugiere utilizar un criterio estadístico. Entre más grande la muestra, el margen de error se reduce. En [Anderson 1999], se recomienda la estimación de la muestra, en poblaciones grandes, mediante la siguiente expresión:

$$n=(N*s^2)/(N*(B^2/4)+s^2)$$

[8] Ídem.

Donde **n** es el tamaño de la muestra, **N** el tamaño de la población, **s** es la desviación estándar de la muestra y **B** es la cota de error deseada de la prueba estadística. **N** es usualmente conocido, **B** es establecido por el que realiza la prueba según la confiabilidad que le quiera dar al diagnóstico, a mayor confiabilidad, el valor de B será más pequeño y por tanto el valor de **n** tenderá a incrementarse.

Una limitación es el cálculo de **s**, que requiere de la existencia de la muestra, cuyo tamaño precisamente se está calculando. En este caso **s** se puede estimar de varias formas, entre ellas, Anderson et al. mencionan: recurrir a un resultado anterior, realizar un muestreo de prueba o estimarla mediante la interpolación de resultados conocidos.

En general, en poblaciones grandes, una muestra aceptable varía entre el 3% y 10% del tamaño de la población. No hay que olvidar que el tamaño de la muestra no debe ser tan pequeño que invalide la prueba, ni tan grande que entorpezca su operacionalización.

Un usuario con conocimiento experto puede ayudar a escoger los elementos de la muestra, velando por que éstos sean significativos. Por ejemplo, si se trata de una cuenta por cobrar, la muestra debe incluir algunas de las cuentas con mayor movimiento y algunas de las que tienen saldos más cuantiosos.

Escribir guión de pruebas

Un guión de pruebas es una guía del procedimiento para realizar las pruebas. Algunos elementos que lo componen son:

- El contexto de la prueba, por ejemplo: fechas de proceso, especificación de la muestra, valores de ambiente del sistema, etc.
- Los datos que se inspeccionan
- Los procesos que se aplicarán
- Los resultados esperados según la definición de cumplimiento
- El procedimiento para verificar si un resultado es el correcto
- El papel (rol) del usuario en las pruebas

El procedimiento de validación de un resultado se especifica mediante frases como estas: “Los datos acerca de saldo e intereses en la impresión del comprobante de pago se verifican mediante la siguiente expresión ....., los resultados se consideran correctos si la diferencia entre el cálculo realizado por el ejecutor de la prueba no difiere con el indicado en el comprobante por más de un 0.5%”.

Preparar ambiente de pruebas

Es importante que el ambiente de pruebas sea independiente del ambiente de operación para evitar interferencias indeseables. Los datos para las pruebas deben estar disponibles así como los criterios para aceptar las pruebas.

Ejecutar pruebas

Se ejecutan las pruebas según lo especifica el guión de pruebas, y se documenta cada resultado.

Cada prueba es evaluada con una calificación normalizada entre 0 y 100, según sea el grado en que se satisfaga la definición de cumplimiento. Luego esta calificación se incorpora a la evaluación final.

2.5 Criterio de usuarios

Es recomendable obtener el criterio de usuarios luego de realizar las anteriores actividades, de esa manera se facilita el evaluar en justa medida los criterios parcializados.

Usuarios operativos

Saben especificar los problemas del día a día en su interacción con las aplicaciones. De esta manera se puede saber dónde la falta de calidad de datos se expone al crear problemas en la calidad del servicio de la aplicación.

Usuarios que toman decisiones

Aún si la falta de calidad no se muestra en el día a día, puede ser que sí lo haga al mostrar información resumida en informes de cierre a final de mes, al generar reportes gerenciales, etc.

2.6 Calificación del diagnóstico de datos

Los elementos anteriormente reseñados pueden ser sujetos de una métrica, ser ponderados y de esa manera estimar cuantitativamente la calidad de los datos.

La siguiente tabla presenta un posible modelo de cuantificación:

Tabla 1 : Evaluación del diagnóstico de calidad de datos

Concepto de Evaluación	**Ponderación**	**Resultado diagnóstico**	**Evaluación**
Aspectos contextuales			
Base de datos			
Inspección de interfaz			
Criterio de usuarios			
Pruebas de diagnóstico			
Total	**100%**	**----------**	

La ponderación es un valor porcentual en el rango de 0 a 100, que especifica el valor que tiene cada concepto dentro de la calidad total de los datos.

La columna resultado diagnóstico representa un valor de 0 a 100 obtenido de las evaluaciones a los ítemes que agrupan cada concepto de evaluación, que se explican más adelante.

La columna Evaluación se calcula de la siguiente manera:

Valor[Evaluación]:=Valor[Resultado diagnóstico]* Valor[Ponderación]/100

Para cada fila en la Tabla 1 debe existir un procedimiento de evaluación que a su vez refiere a otra tabla, de donde se toma el resultado para ser indicado en la columna con título *Resultado del diagnóstico*.

Para el caso del concepto de evaluación llamado Aspectos Contextuales se propone una tabla con la siguiente estructura:

Tabla 2: Evaluación del Diseño de Bases de Datos

Ítem de evaluación	**Métrica**	**Ponderación**	**Resultado del diagnóstico del ítem**	**Resultado ponderado**
Normalización no controlada				
Cantidad de atributos nulos				
Cantidad de atributos derivados				
Integridad referencial				
.....				
Total	**---------**	**100%**	**----------**	

La columna métrica contiene una descripción del procedimiento para evaluar cada ítem. El resultado debe ser un valor entre 1 y 100. Las métricas deben ser procedimientos sencillos y rápidos de implementar. Por ejemplo, la siguiente tabla operacionaliza una métrica para evaluar integridad referencial:

Tabla 3: Métrica para evaluar integridad referencial

Criterio	Valor
Todos los atributos que refieren a una llave en otra entidad (tabla) son declarados como llave foránea	100
Menos de un 5% de estos atributos no son declarados	85
Entre un 5% y un 20% no son declarados	60
Más de un 20% no son declarados	30

Otra métrica ayuda a evaluar corrección en atributos alfanuméricos; por ejemplo, un atributo que almacena las señas de una dirección de cobro:

Tabla 4: Métrica para evaluar señas de dirección de cobro

Criterio	Valor
El correo no devuelve ningún envío de cobro	100
El correo devuelve entre 1 y 20 envíos por mes	90
....	

La métrica para evaluar el criterio de usuarios puede estar basada en encuestas de opinión.

Es inevitable que algunas métricas reflejen algún grado de subjetividad, sobre todo al evaluar aspectos contextuales. En este caso, es recomendable que la ponderación sea baja, para reducir su distorsión en la evaluación final.

La columna *Resultado Ponderado* se calcula así:

Valor[Resultado ponderado]:=Valor[Ponderación]*Valor[Resultado diagnóstico]/100

Para determinar el valor ponderado de cada ítem se puede recurrir al criterio colegiado de usuarios y de técnicos mediante la realización de un taller de trabajo.

2.7 Evaluación del diagnóstico

Del análisis de los resultados encontrados, es importante encontrar las causas de los problemas en los datos y documentarlas. Sin ser exhaustivo, entre las posibles causas se enumeran las siguientes:

- Procesos semi-manuales
- Datos redundantes
- Interfaces discordantes entre sistemas
- Ausencia de validaciones
- Interfaces de usuario inducen a error
- Mala capacitación de usuarios
- Lagunas de seguridad o integridad
- Instrumentos de captura de datos (pe: formularios) inadecuados
- Problemas de corrección del sistema

3 Aspectos por considerar

A continuación se presentan consideraciones que conviene tomar en cuenta al emprender diversos tipos de proyectos informáticos:

- Adquisición de sistemas
- Desarrollo de sistemas
- Migración de datos

3.1 Consideraciones al gestionar una adquisición de sistemas

Al diseñar el proceso de adquisición de un sistema, debe evaluarse la calidad de datos que éste produce con base en un diagnóstico que puede adaptarse del indicado en el capítulo anterior.

Ejemplo de algunas adaptaciones, son las siguientes:

- El concepto de "Aspectos Contextuales" puede incluir los siguientes incisos:
  - Madurez del producto (medido en meses de operación del sistema)
  - Uso de estándares en la implementación de códigos (p.e. uso del estándar internacional para codificar código de ocupación de un cliente)
  - Uso de estándares internacionales en procesos (p.e. el estándar FASB 52 para registro contable multimoneda).
  - La plataforma es basada en un SABD relacional.
- En lugar del concepto "Criterio de usuarios", podría utilizarse: "Criterio de otros clientes".
- El concepto: "Pruebas de diagnóstico" se puede realizar con base en simulaciones.

3.2 Consideraciones al desarrollar sistemas

La calidad de datos es un valor agregado de un buen diseño de sistemas, tanto en el nivel de la estructura de los datos como en los procesos. A continuación se describen algunos lineamientos que compelen el aseguramiento de la calidad de los datos.

Usar métodos de orientación a objetos para diseñar bases de datos

Al diseñar la base de datos, encapsular la definición de la estructura de información con los procesos que la acceden, de manera que las reglas de calidad se incorporen como parte de la descripción de los datos mismos.

Características de un buen diseño de bases de datos

Algunos de los elementos que aseguran un buen diseño ya fueron comentados anteriormente en la sección 2.2. Estos elementos se enumeran aquí, con algún comentario adicional en caso necesario.

- Normalizado o con desnormalización controlada
- Selección conveniente de llaves
- Integridad referencial
- Uso de reglas de integridad
- Valores nulos
- Atributos derivados
- Reducir la representación de estados

  Un efecto de reducir la cantidad de atributos derivados es la disminución de la representación de estados de la base de datos en un momento en el tiempo. Esta guía de diseño reduce complejidad, mejora la reversabilidad de las actualizaciones y por lo tanto, incrementa la calidad de los datos al minimizar errores de consistencia debidos a "deshacer" incompletos de actualizaciones inadecuadas.

  Un ejemplo es este: es común en los sistemas de contabilidad almacenar los saldos de las cuentas al realizar el cierre de un período contable. Se puede obviar este almacenamiento si la información base para el cálculo de los saldos al cierre está siempre disponible y accesible rápidamente. Se reduce complejidad pues ya no es necesario representar los cierres (estados de los saldos de las cuentas en un momento en el tiempo), aunque es necesaria una mayor cantidad de procesamiento en el recálculo de los saldos cuanto tal información es requerida.

Procesos de conciliación

Es conveniente que el diseño de un sistema incluya un módulo de conciliación. Las datos de un sistema se concilian con respecto de los datos que provee un sistema independiente. Por ejemplo, las sumas de control de un Sistema de Cuentas por Cobrar deben conciliarse con los datos sobre ingresos que ofrece el Sistema de Tesorería.

Para el siguiente ejemplo, se hacen las siguientes suposiciones acerca de Sistemas de Cuentas por Cobrar:

- El sistema afecta una cuenta por intereses y otra por amortización.
- Los cierres para procesos de consolidación se realizan diariamente.
- El registro de movimientos permite el ingreso de movimientos para la realización de ajustes.

Para la implementación de un proceso de conciliación es necesaria una entidad que represente los siguientes atributos:

- *Fecha de proceso*
- *Suma de ingresos de tesorería por cuentas por cobrar*
- *Suma de amortizaciones aplicadas*[9]
- *Suma de intereses aplicados*
- *Diferencia*
- *Ajustes realizados*

El atributo diferencia se calcula así:

Diferencia := *Suma de ingresos de tesorería por cuentas por cobrar* -
(*Suma de amortizaciones aplicadas* +
Suma de intereses aplicados) +
Ajustes realizados

El atributo *ajustes realizados*, acumula el monto de los movimientos por ajustes (créditos o débitos) a la fecha de proceso. Un ajuste se crea cuando se detecta que la diferencia no es cero, el monto del ajuste coerce a que la diferencia tienda a cero[10]; una diferencia en cero significa que el movimiento ha sido conciliado en esa fecha.

3.3 Consideraciones al migrar datos

La migración de datos se realiza por alguna de las siguientes razones:

- Implantación de nuevo sistema
- Escalamiento de plataforma
- Escalamiento de la aplicación
- Re-estructuración de datos[11]

La migración debe asegurar que la calidad de los datos en su nuevo contexto es la esperada, mediante la ejecución de las actividades que seguidamente se describen.

Mapeo de datos de la vieja plataforma a la nueva

El mapeo (correspondencia) de los datos de la plataforma actual a la que se migra constituye una herramienta para:

- Validar la compatibilidad de las reglas de validación de atributos, para cada par de atributos (atributo en platafoma actual, atributo equivalente en plataforma por adquirir). Esto es importante para normalizar el concepto de calidad de datos entre la plataforma actual y la nueva.

[9] Este atributo y los dos siguientes son atributos derivados que pueden calcularse del registro de movimientos de una cuenta x cobrar, al ser derivados puede prescindirse de ellos.
[10] Siempre que el movimiento por ajuste pueda ser justificado.
[11] En atención a nuevos requerimientos para corregir deficiencias de rendimiento (en la manipulación o acceso a los datos) o para facilitar modificaciones futuras del sistema.

- Diseñar los instrumentos para diagnosticar la calidad de los datos en la plataforma actual.

Diagnóstico de la calidad de datos

Se realiza un diagnóstico de la calidad de datos, como se describe en el capítulo 2.

Diseñar la carga de datos y especificar las condiciones de aceptación para iniciar la operación

El análisis de los resultados del diagnóstico de datos facilita la especificación de la carga de datos así como los recursos para su implementación.

Un elemento que provee un contexto para este diseño es la definición de cumplimiento de calidad de los datos que se migran para que sean aceptados por el nuevo sistema. Un ejemplo de una definición de cumplimiento con respecto del atributo saldo es el siguiente: "el 98% de los saldos en las cuentas por cobrar que se migran cumplen que el saldo = sumatoria (créditos) – sumatoria (débitos)".

Para cada definición de cumplimiento, se debe especificar el instrumento para validarla, mediante un informe que despliegue cifras tales como totales y porcentajes, que permitan certificar la calidad a los funcionarios que realizan o auditan la carga de datos.

Documentar la carga de datos

La carga de datos se documenta con los informes que respaldan la satisfacción de las definiciones de cumplimiento, firmados por los actores encargados de la carga de datos.

4 Testimonio de experiencias

4.1 Elaboración de un modelo de datos corporativo

El proyecto se realizó para un empresa grande, cuyos productos o servicios son variados. Cada servicio es llevado a cabo por divisiones organizacionales que se han desarrollado con un cierto nivel de independencia, y han llegado incluso a gestionar sus propios sistemas de información.

La empresa es grande, con presencia en las principales ciudades del país. Para hacer frente a la competencia y poder comparar su gestión con las condiciones socioeconómicas de cada región, la empresa necesita un modelo de datos corporativo con el fin estandarizar la estructura de la información tal que permita consultar datos de gestión global.

Al hacer el diagnóstico de datos, que fue la base para proponer acciones para construir el modelo, se encontraron las siguientes situaciones:

- Los datos de un mismo cliente se encuentran definidos muchas veces en los sistemas que implementan los diferentes procesos de la empresa.
- La información de un mismo cliente no es consistente, debido a que los formatos de campos que refieren a un mismo concepto son diferentes o porque la enumeración de códigos no coincide.
- Conceptos equivalentes tienen nombres distintos en diferentes sistemas, creando un vocabulario inconsistente entre los usuarios.
- Las enumeraciones de códigos para conceptos equivalentes son diferentes, así que las interfaces entre sistemas incluyen homologaciones de códigos.

El modelo de datos fue utilizado como un criterio de selección de un sistema corporativo y fue útil para comparar qué tanto la arquitectura de los datos de un sistema candidato es compatible con lo requerido por la empresa.

El diagnóstico fue útil en la estimación de los recursos requeridos para reparar datos, estandarizar formatos y enumeraciones de códigos, realizar carga de datos y interfazar aplicaciones.

Un valor agregado no cuantificable del diagnóstico fue el de tomar conciencia acerca de los riesgos de la organización al evidenciarse el estado de la calidad de sus datos.

4.2 Desarrollo de un sistema

En el desarrollo de un sistema de información se encontraron los siguientes indicadores de una pobre calidad de datos:

- Las cuentas contables afectadas por el sistema tenían entre 1 y 5 años de no estar conciliadas.

- Existe un anterior sistema que no satisface los requerimientos mínimos, entre ellos: un 20% de las operaciones se procesan manualmente.
- Algunos de los datos provienen de interfaces con otros sistemas que tienen una pobre calidad de datos.
- Cuando el desarrollo del proyecto se encontraba en estado muy avanzado (75%) se encontró que el usuario principal (el jefe del Departamento) realizaba actividades fraudulentas al amparo de la anarquía en el manejo de información.

Las consecuencias directas de estos hechos fueron las siguientes:

- Una reevaluación de toda la información aportada por el usuario principal, y a partir de ello la realización de una reingeniería al sistema en desarrollo.
- Atraso muy costoso en la implantación del sistema mientras se incrementaba la calidad de los datos a un nivel aceptable. Esto incluyó la cedulación de los clientes (40,000 clientes) y la conciliación de las cuentas.
- El desarrollo de un proceso de carga de datos, que en sí mismo era un proyecto de tal complejidad que requirió un esfuerzo significativo de toda la organización.

El proyecto quedó en operación gracias al esfuerzo heroico de muchas personas. Sin embargo, nunca fue posible incrementar la calidad de datos a un nivel establecido como deseable; esto ha afectado el rendimiento del sistema desde su inicio de operaciones.

Un diagnóstico de la calidad de los datos al *inicio* del proyecto habría puesto en evidencia los riesgos de su desarrollo, así se habrían tomado medidas correctivas a tiempo.

4.3 Un proyecto del Y2K

Algunas de las técnicas expuestas en este documento fueron aplicadas para diagnosticar calidad de datos enfocada a los requerimientos en el manejo de fechas ante el cambio de milenio.

Se encontró una alta correlación entre los sistemas de alto riesgo y aquellos con pobre calidad de los datos. Un diagnóstico basado en la inspección de la base de datos, enfocado en formatos y datos correlacionados con fechas, fue muy útil para establecer riesgos que sustentaran el establecimiento de prioridades para las actividades de reparación y certificación.

5 Conclusiones y recomendaciones

La calidad de los datos es un valor agregado de la calidad en el procesamiento de la información. La calidad de los datos es un espejo de la calidad de la organización en la gestión de información.

Un diagnóstico puede ayudar a definir estrategias que reduzcan riesgos en la realización de proyectos informáticos tales como: desarrollo de sistemas, evaluación de sistemas previa a una adquisición, diagnóstico de sistemas, migración de datos, mantenimiento e integración de sistemas.

Un *diagnóstico general* que dé pistas acerca de calidad de los datos puede realizarse rápidamente mediante la inspección de tres elementos: aspectos contextuales, diseño de base de datos y criterios de usuarios. El cuarto elemento, realización de pruebas de diagnóstico, requiere mayores recursos. Este elemento confirmará lo encontrado mediante la primera inspección y aislará los síntomas.

Un *diagnóstico detallado* debe incluir la inspección exhaustiva de la base de datos y la realización de pruebas. Estas dos actividades requerirán mayores recursos, y su objetivo consiste en confirmar lo encontrado mediante un diagnóstico general y aislar el origen de cada síntoma.

5.1 Recomendaciones

Es recomendable aplicar un diagnóstico de calidad de datos en el proceso de gestión de un proyecto informático con el fin de determinar riesgos. El diagnóstico puede ser general o detallado dependiendo de la importancia del proyecto o de su sensibilidad a los riesgos.

Una organización debe aplicar periódicamente diagnósticos de la calidad de los datos, ya sea generales o detallados, dependiendo de los eventos recientes que puedan afectar el desempeño de sistemas de información en operación, por ejemplo: ajustes debidos a nuevos requerimientos, integración, etc. Detectar a tiempo la causa de problemas en la calidad de datos reduce riesgos y costos.

Inculcar el velar por la calidad de datos como un valor de la empresa, más allá de un contexto informático, es una manera de compeler a la organización a asumir un cultura de calidad en todo sentido, por ejemplo: calidad del servicio al cliente, calidad del producto terminado, etc.

En organizaciones que desarrollaron proyectos de reparación y certificación de sistemas para el año 2000, es una buena acción evaluativa el realizar un diagnóstico de la calidad de los datos. Uno de los beneficios sería el detectar posible corrupción de datos en bases de datos activas o históricas, debida a procesos de reparación sobre las estructuras de datos o el código de los programas. En el anexo se adjunta un artículo, escrito por Ed Yourdon [Yourdon 1999], que llama la atención sobre este asunto.

Anexo

Corrupción de la información: el asesino silencioso del error del año 2000

Ed Yourdon[12]

Siempre que pensamos en el error del año 2000, tenemos la tendencia a concentrarnos en los problemas "visibles", por ejemplo, una falla incrustada en el sistema que hace que una refinería explote, es algo que hay que arreglar en el momento de la falla. Mientras tanto, existe otro problema del año 2000 que es mucho más insidioso y requerirá que se le dedique atención y recursos durante todo el próximo año: el problema de la corrupción de datos. No me refiero a una corrupción masiva, repentina y visible de información – como un sistema de planillas que empieza a hacer cosas extrañas y fija el salario de los empleados en cero.

Lo que me preocupa es el error del año 2000 que dañará sólo un pequeño porcentaje de una base de datos, de tal forma que su impacto no será visible de manera inmediata. Por ejemplo, ¿qué ocurriría si un error actualizara el registro de una base de datos activa (actual) correctamente, pero también dañara una porción pequeña del registro de una base de datos inactiva (histórica) – como un corrector de códigos que reemplaza de manera correcta el campo de "año" de dos dígitos ("AA") en el registro de una base de datos activa, por un campo de cuatro dígitos ("AAAA"), pero contiene un error que altera los primeros dos bytes de un registro adyacente? Podrían transcurrir meses o años antes que el registro de una base de datos inactiva sea accedido, o suficientes registros inactivos hayan sido dañados, lo cual provocaría el colapso de toda la base de datos. Y el problema puede ser aún más sutil si el error abarca interfaces entre sistemas operados por diferentes organizaciones.

La corrupción de datos no es un concepto nuevo y no es característico sólo del año 2000. Pero, irónicamente, algunas organizaciones no tomaron conciencia de los problemas de corrupción de datos a largo plazo en sus bases de datos, hasta que empezaron a trabajar en soluciones para el error del año 2000.

Entonces, ¿cómo enfrentamos la corrupción de datos? La mayoría de las organizaciones creen que pueden evitar el problema mediante pruebas rigurosas y cualquier mecanismo de control de errores que se encuentre incorporado al código de la aplicación y al paquete del sistema de administración de bases de datos (DBMS) del vendedor. Pero podrían estarse engañando a ellas mismas: las probabilidades de evitar la corrupción en una base de datos con 10 millones de registros, la cual ha estado en operación durante 10 años son pocas. En realidad, es posible que la única razón por la que las organizaciones tienen sistemas estables, es porque los fueron construyendo poco a poco y los modificaron relativamente despacio a lo largo del tiempo.

[12] Yourdon dirige el servicio del año 2000 en Cutter Consortium en Arlington, Mass. Pueden contactarlo en yourdon@acm.org. Este artículo apareció en Computer World América Central 3(21), diciembre de 1999.

El año 2000 es básicamente diferente porque implica realizar cambios masivos en todos los sistemas a la vez. Sí, las pruebas han sido amplias en la mayoría de las organizaciones grandes y probablemente eliminaremos todos o casi todos los errores visibles. Pero se necesita un enorme optimismo para presumir que habremos eliminado los errores sutiles que provocan los problemas insidiosos de corrupción de información – especialmente cuando vendedores independientes de verificación y validación como Cap Gemini, MatriDigm y Reasoning Systems, informan que encuentran entre 400 y 900 errores por cada millón de líneas de código que supuestamente habían sido solucionados y probados.

Creo que es más realista suponer que los problemas de corrupción de información ocurrirán y que podríamos no verlos durante meses o años después del primero de enero. Así que continuamos preguntándonos: ¿Cómo enfrentamos la corrupción de datos? La solución es simple y obvia, aunque para nada infalible. Es necesario desarrollar amplios programas que puedan auditar, verificar y controlar la integridad de los datos y aplicarlos periódicamente durante el año 2000 e incluso más allá. Estos programas deben ser aplicados diariamente o al menos semanalmente durante los primeros meses, dependiendo del tamaño de la base de datos y de la cantidad de ciclos extra disponibles en el procesador. Después, según los resultados obtenidos, podremos relajarnos y aplicar los programas mensualmente.

Existen para la venta paquetes para la verificación de datos y algunas organizaciones han desarrollado sus propios programas para minimizar la corrupción de la información. No obstante, esto no es algo que se hace comúnmente y la mayor parte de las organizaciones que visito no han planeado gastar dinero o recursos informáticos en esta clase de estrategia durante el próximo año. Creo que este descuido les saldrá caro y que llegará a agravar el problema del año 2000 más de lo que debiera.

Bibliografía

[Anderson 1999] Anderson, David; Sweeney, Dennis; Williams, Thomas. Estadística para Administración y Economía. International Thompson Editores. Sétima Edición. 1999.

[IEEE 1993] IEEE 1062-1993. *Recommended Practice for Software Acquisition.*

[Muñoz 2000] Muñoz, Lilia. Calidad de los datos: Un enfoque conceptual. Club de Investigación Tecnológica, febrero 2000.

[Yourdon 1999] Corrupción de la información: el asesino silencioso del error del año 2000. Computer World América Central, diciembre 1999.