Aplicando un enfoque de Ingeniería de Software a la Calidad de Datos

Mónica Bobrowski
Daniel Yankelevich
José Jiménez Arlegui
Practia Consulting

La CRITICIDAD de la Calidad de los Datos

- Importancia creciente del “valor de la información
- Información en el sistema = información real
  - El sistema ES la realidad
- Las decisiones tomadas en base a datos de mala calidad tienen un alto impacto económico

La NECESIDAD de la Calidad de los Datos

"Las decisiones no pueden ser mejores que los datos en que se basan"

- Perspectiva del cliente
  - ¿Puedo confiar en una compañía que maneja datos erróneos?
- Perspectiva de la compañía
  - Importancia de contar con datos confiables, de calidad, a la hora de tomar decisiones
  - Tener datos de mala calidad conlleva a la pérdida de clientes, dinero, confianza
  - Necesidad concreta de datos de calidad en emprendimientos de business intelligence, data warehousing, data mining, knowledge management

El PROBLEMA de la Calidad de los Datos

"El sistema funciona perfectamente. Obviamente, si se cargan datos erróneos, ¿qué puede hacer el sistema?"

- Los buenos sistemas de software ayudan pero no garantizan datos de calidad
- Los buenos usuarios ayudan pero no garantizan datos de calidad
- No se trata de "limpiar" datos periódicamente, se trata de evitar tener datos "sucios". Esto es más barato desde todo punto de vista.

Dimensiones de la calidad de los datos

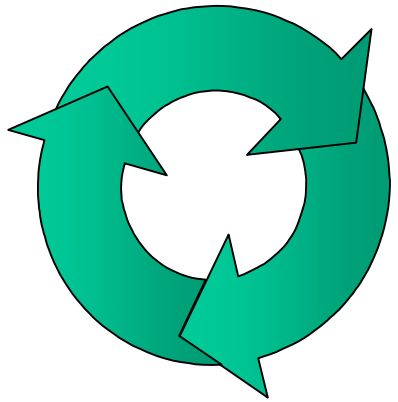
- Datos de calidad no significa
  DATOS PERFECTOS
- Sino que es un concepto relativo; se trata con dimensiones:
  - completitud
  - relevancia
  - consistencia
  - confiabilidad
  - vigencia
  - corrección
  - precisión
  - concisión
  - ...

El enfoque hacia la Calidad de Datos

- Usualmente:
  - Sólo se corrige el estado actual de los datos
  - Se implementan soluciones ad-hoc
  - Se apela sólo a técnicas de “data cleansing” (limpieza de datos) aplicandolas a problemas de filiación y domiciliación
  - No se trabaja sobre el modelo de datos y no se apunta a la fuente de generación del dato
  - No se analizan ni corrigen los procesos que llevan a producir datos erróneos
- La propuesta:
  - **Aplicación de técnicas de ingeniería del software al problema de calidad de datos**

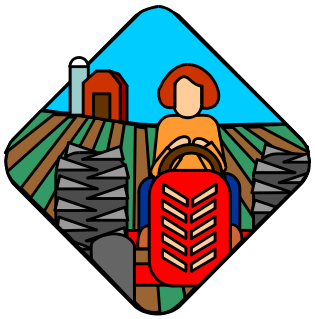
Metodología

- Etapa 1. Identificación, Medición, Diagnóstico y Plan de Mejoras
- Etapa 2. Tratamiento de los Datos
  - Tratamiento Correctivo
  - Tratamiento Preventivo
- Etapa 3. Mantenimiento de la Calidad de los Datos



Herramientas y técnicas utilizadas

- **De análisis:** Permiten analizar sintácticamente la información
- **De detección de reglas de negocios:** Permiten analizar contenidos de campos para descubrir patrones en los datos y relaciones entre los datos
- **De limpieza:** Permiten corregir datos. Manejan extracción, estandarización, matching, consolidación de datos duplicados, transformación de tipos de datos, etc.
- **GQM:** Por Goal Question Metric (métricas orientadas por preguntas), es una técnica para elaborar métricas a partir de la identificación de problemas con usuarios y técnicos
- **Templates y Checklists:** Plantillas predefinidas para la especificación de los objetivos de calidad y la documentación ulterior de los resultados obtenidos



¿Qué quieren los usuarios?

- A lo largo de nuestra experiencia hemos recibido de ellos las siguientes demandas:
  - Datos correctos, consistentes y completos
  - Datos relevantes
  - Datos vigentes
  - Visualización de los datos de una manera apropiada en el marco de las aplicaciones utilizadas
  - Fácil acceso a los datos
  - Datos seguros

Problemas de Calidad encontrados

- **Asociados a la instancia**
  - datos que han cambiado en el mundo real, y que no fueron actualizados
  - datos que provienen de distintas fuentes, deberían ser consistentes y sin embargo no lo son (por ejemplo, catálogos de servicios o productos en diferentes aplicaciones que no se mantienen integrados)
  - datos que no han sido almacenados con la precisión necesaria (por ejemplo, Y2K)

Problemas de Calidad encontrados (2)

- **Asociados al modelo de datos**
  - si se detecta que hay información que no está presente porque no hay forma de almacenarla, entonces el modelo de datos físico está incompleto
  - el mundo que se quiere representar evolucionó, no se tradujeron los cambios al modelo, entonces el modelo perdió vigencia
  - reglas de negocio no trasladadas al modelo
  - inconsistencia entre modelo lógico y modelo físico

Problemas de Calidad encontrados (3)

- **Asociados a los Procesos**
  - distintas personas cargan la misma información haciendo distintas asunciones
  - se carga con una asunción y se usa con otras
  - modificaciones manuales-por procesos
  - gente que hace modificaciones pero no debería estar autorizada para hacerlas
- **Asociados a errores de software**
  - datos obligatorios que no se asumen como tales y por lo tanto no se cargan
  - interfaces poco amigables
  - rangos de valores que no se respetan (la producción acumulada crece o se mantiene, pero nunca decrece)

Aspectos de la calidad a atacar

- Se debe trabajar sobre la instancia
- Se debe trabajar sobre el modelo de datos
- Se debe trabajar sobre los procesos que intervienen en la generación y modificación del dato

Experiencia I

- Situación: Banco con datos de clientes provenientes de diversas fuentes. Problemas en los datos de los reportes enviados al Banco Central. Realización de trabajos periódicos de depuración de datos.
- En un mes se realizo un diagnóstico que detectaba los principales problemas de los datos y sus causas, y un plan de mejoras preventivas.
- Antes de llevar a cabo el plan, se volvieron a ejecutar las métricas implementadas para el diagnóstico y se encontró un aumento importante en la cantidad de datos erróneos.

Experiencia II

- Situación: Empresa petrolera. Problemas de carga de información histórica a pedido del ente regulador. Análisis de completitud del modelo y de consistencia de la información.
- En un mes y medio se elaboró un diagnóstico que detectaba los principales problemas del sistema, de los datos y sus causas, y un plan de mejoras preventivas.
- Se realizaron modificaciones al sistema para que el mismo se adaptara a las necesidades. Se disminuyó el volumen de inconsistencias del sistema.

Experiencia III

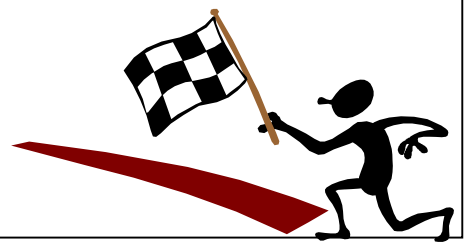
- Situación: Sistema de información de una red de transmisión de datos. Modelo de datos inexistente. Problemas con la calidad de los datos.

- En dos meses se generó el modelo de datos, se reportaron mediciones de problemas de consistencia, precisión, vigencia y de procedimiento (p.ej. carga errónea de circuitos), y se formularon recomendaciones de mejora, entre ellas cómo cargar los datos y cómo estandarizar la codificación.

- Este trabajo venía siendo desarrollado desde hacía varios meses sin resultados visibles y sin una metodología adecuada.

Resumiendo...

- Aplicación de técnicas de Ingeniería del Software
- Focalización en la visión del usuario
- Se trabaja sobre la instancia (contenido) y el modelo de datos
- Se atacan los procesos que intervienen en la generación y modificación del dato
- Se implementan soluciones sistemáticas
- Se define la Calidad Esperada para cada caso
- La solución se instancia al mercado vertical correspondiente



Conclusiones

"Un hombre con un reloj sabe qué hora es; un hombre con dos relojes nunca está seguro",
Mark Twain

- Los negocios dependen cada vez más de la calidad de los datos con que cuentan
- Necesidad de focalización en el usuario
- Datos de calidad en términos de los objetivos de la organización, no datos perfectos
- No alcanza con buen software
- Conveniencia de soluciones sistemáticas, probadas, preventivas
- El dato es su instancia y su modelo