DATA QUALITY

Data quality se refiere a los procesos, técnicas, algoritmos y operaciones destinadas a mejorar la calidad de los datos.

El problema aparece cuando: los datos no significan lo que deberían, muchas fuentes, distintos datos sobre la misma información, problemas costosos y con gran repercusión, cambios sistemáticos externos al proceso de grabado (Data Glitches)

Data Quality Continuo:

Los datos y la información no son estáticos se engloban en una colección de datos y siguen un proceso de uso:

- Recopilación de datos (Data gathering)
- Entrega de datos (Data delivery)
- Almacenamiento de datos (Data storage)
- Integración de datos (Data integration)
- Recuperación de datos (Data retrieval)
- Data mining/análisis

Recopilación del dato (gathering):

- Problemas: Entrada manual, sin estándar en los formatos y contenidos, varias fuentes de entrada del mismo dato (lo que genera duplicados), aproximaciones, sustituciones, restricciones, errores de medición.
- Soluciones Potenciales:
  1. Preventivas: diseñar una arquitectura del proceso, gestión de procesos
  2. Paliativos: enfoques de limpieza, enfoque de diagnósticos

Entrega de datos (Delivery): Definir protocolos de transmisión seguros, verificación, relaciones, acuerdos de interfaz.

Soluciones:

1. Preventivas: Diseñar una arquitectura del proceso y gestión de procesos
2. Paliativos: Enfoques de limpieza y enfoque de diagnósticos

Almacenamiento de datos (Storage)

- Recibes un dato
- Problemas en el almacenamiento físico
- Almacenamiento lógico (Modelo Entidad Relación):
  - Metadatos pobres
  - Modelos de los datos incorrectos
  - Modificación ad-hoc
  - Restricciones de hardware/software
- Soluciones:
  - Metadato → Documentar y publicar las especificaciones de los datos
  - Planificación → asumir que todo lo que pueda salir mal pasara
  - Exploración de los datos (antes de almacenar):

Usar herramientas de exploración y minería de datos para validar: especificaciones que se asumieron previamente, comprobar si algo ha cambiado

Integración de datos (Data Integration):

- Problemas:
  - Heterogeneidad de los datos: claves no comunes, diferencias en el formato de los campos
  - Definiciones diferentes → ¿Qué es un cliente?
  - Sincronización temporal
  - Datos heredados
  - Factores sociológicos
- Soluciones:
  - Herramientas comerciales y librerias para coincidencias parciales
  - Decisiones estructurales en la compañía
  - Visualización y exploración de datos

Recuperación de los datos (Data Retrieval):

En ocasiones los datos exportados son vistas de los datos actuales. Lo que provoca los siguientes problemas:

- Fuentes (originales) no entendidas adecuadamente
- Utilización de datos derivados no entendidos
- Simplemente errores (inner join vs outer join y valores nulos)
    - Restricciones computacionales
    - Incompatibilidades

Data Mining and Analysis:

- Problemas del análisis:
  - Escalado y rendimiento
  - ¿Bandas de confianza?
  - Cajas negras y diagramas de validación frente a la aleatoriedad
  - Adjunto a los modelos
  - Experiencia de dominio suficiente
  - Empirismo casual
- Soluciones:
  - Exploración de los datos: Determina qué modelos o técnicas son apropiadas, encuentra bugs en los datos y mejora tu conocimiento del dominio
  - Validación constante
  - Responsabilidad: ciclo de mejora

DATA QUALITY 2

Hay muchos tipos de datos, que tienen diferentes usos con sus problemas de calidad característicos:

- Datos federados*
- Datos de alta dimensionalidad
- Datos descriptivos
- Datos longitudinales
- Datos en streaming
- Datos de web (scrapeados)

usos de los datos: Operativos, análisis de datos agregados, relación entre clientes, etc.
Integración de datos: Los datos son inútiles si no se conocen las *reglas* que hay detrás

Idoneidad de los datos: ¿Qué respuestas se pueden obtener de los datos?: Uso de datos base*, faltan datos relevantes

Limitaciones:

- Hay limitaciones estáticas debidas a los esquemas
- Hay otras limitaciones dinámicas que tienen que ver con el flujo de procesado
- Se sigue la regla de 80 – 20

METRICAS DE DATA QUALITY

- Queremos una medida cuantitativa
  - Que indique que está mal y cómo se puede mejorar
  - Hay que tener en cuenta que no hay un conjunto de datos perfectos, las medidas nunca serán óptimas
- Tipo de métricas
  - Estáticas frente a dinámicas
  - Operativas frente a diagnósticos
- Las métricas deben indicar una dirección de mejora el uso de los datos
- Un número muy elevada de métricas es posible (optimas)

Ejemplo de métricas de data quality:

- Conformidad con el esquema
- Conformidad con las reglas comerciales
- Exactitud (Accuracy)
- Accesibilidad
- Interpretabilidad
- Glitches (problemas técnicos) en el análisis
- Finalización exitosa del proceso de extremo a extremo

HERRAMIENTAS TÉCNICAS:

- Necesitamos un enfoque multidisciplinar para solucionar estos problemas
- Gestión del proceso
- Estadísticas
- Base de datos
- Metadata / conocimiento del dominio

Gestión del proceso:

Procesos de negocio que fomentan la calidad del dato:

- Asignar costes a los problemas de calidad
- Estandarización de contenido y formato
- Introducir los datos una vez, hacerlo de forma correcta (incentivar ventas y cuidado de clientes)
- Automatización

- Asignar responsabilidad: ADMINISTRADOR DE DATOS
- End-to-end auditorías de datos y revisiones
- Monitorización de datos
- Publicación de los datos
- Ciclos de feedbacks

Feedback loops:

- En ocasiones lo sistemas de procesamiento de datos se interpretan como sistemas de circuito abierto:
  - Se realiza el procesamiento y se colocan en donde se puedan utilizar
  - Como si los sistemas definidos no pudieran ser erróneos, pudiéramos contemplar el dominio completo de errores y los sistemas no cometieran errores.
- Análogamente a los sistemas de control: bucles de retroalimentación: Es necesario trabajar con ciclos para que el sistema se automejore. Esto son pasos generales de un sistema de control lineal simple que en nuestro caso tiene una mayor complejidad, pero se puede simplificar de esta manera.

Ejemplo:

- Ventas, aprovisionamiento y facturación para un servicio de telecomunicaciones
- La transición entre los distintos departamentos de la organización es una causa común de problemas
- Loops de feedback natural (El cliente se queja si la factura es elevada)
- Bucles de feedback inexistentes (No hay quejas si tenemos un error por defecto)

Monitoring:

La monitorización del proceso y la revisión de los datos son los cimientos del ciclo de feedback del sistema

- Metodos:
  - Seguimiento / auditoría de datos
  - Composición de bases de datos actualizadas incrementalmente con fuentes origen
  - Coherencia obligatoria con una base de datos de registros o metadatos (DBOR)
  - Sincronización del sistema de feedback
  - Publicación de datos y del proceso realizado (para su crítica/uso)

Democratizar el Dato:

- Hacer que el contenido de una base de datos esté disponible de manera sencilla y accesible:
  - Interfaz web (cliente universal)
  - Eliminación de datos: publicar agregados, cubos, muestras, representaciones paramétricas
  - Publicar los metadatos
  - Documentar los procesos realizados sobre los datos, todo el flujo
- Cerrar los circuitos de feedback involucrando al mayor número de personas posible.
- Sorprendentemente difícil a veces.
  - Límites organizacionales, pérdida de control interpretada como pérdida de poder, deseo de ocultar problemas.

Enfoque Estadísticos:

- No utilizan métodos específicos de DQ:
  - En el ámbito estadístico en general los datos son recogidos para el experimento de manera que están muy enfocados al análisis por lo que no hay grandes dificultades
  - Sin embargo, hay métodos para encontrar anomalías y reparar los datos
  - Estos métodos se pueden adaptar a Data Quality
- Cuatro grandes métodos que pueden ser adaptados a data quality
  - Datos perdidos, incompletos, ambiguos o dañados. Ejemplos: censurados, truncados
  - Datos sospechosos o anormales. Outliers
  - Prueba de salida de los modelos
  - Bondad de ajuste