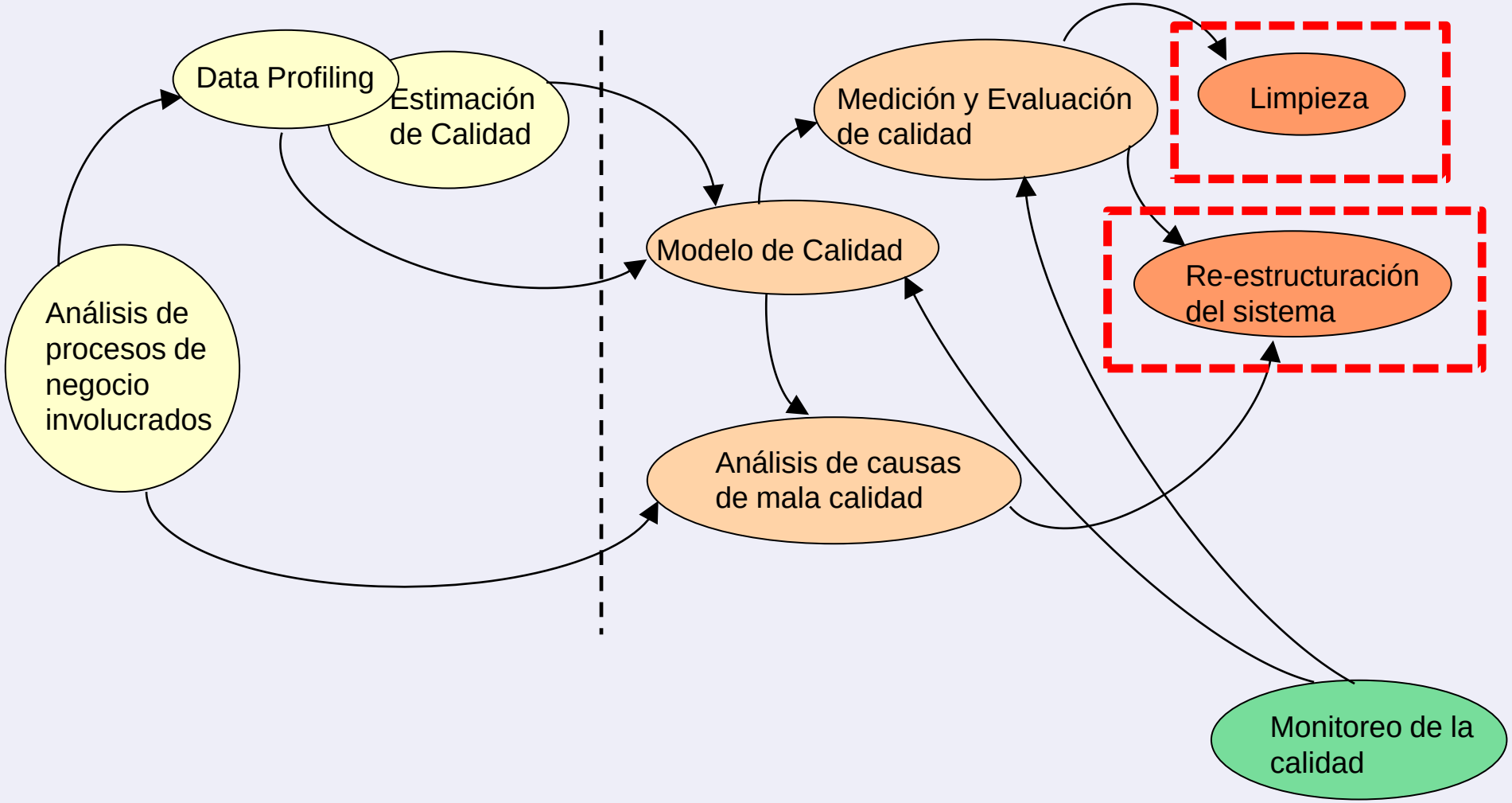
Calidad de Datos e Información

Mejora de Calidad de Datos

Gestión de la calidad en SI



Limpieza de datos

- Identificar y eliminar inconsistencias, discrepancias y errores en datos, para mejorar la calidad
- “data cleaning”, “data cleansing”, “data scrubbing“
- En Data Warehousing
  - **Como parte del proceso ETL (extracción, transformación y carga)**
  - **Hasta un 80% del costo en proyectos de DW**
- En sistemas de integración de datos
  - **“on the fly” para datos integrados virtualmente**
  - **A veces requiere materialización**

Prevención de errores

- Análisis de Causas y Modificación de Procesos

- Localización (o detección) y corrección de errores *no previenen errores futuros*.
  - Ej.: Suponer que un proceso crea o reemplaza 1000 registros nuevos o existentes cada día, cada registro tiene 20 campos y la tasa de errores del proceso es 2%. *400 nuevos errores se producen por día*. A fin de año se habrán producido *140000 errores*.

    → Enorme tarea de limpieza

- Se busca
  - identificar causas (*root-causes*) de los errores
  - eliminar esas causas
  - asegurar que se mantendrá esa ganancia

Plan de Mejora de la Calidad

- Guiado por
  - Modelo de Calidad
  - Metadatos obtenidos de la medición
- Ordenado por
  - Dimensiones y factores
    - Limpieza
    - Eliminación de Causas (modificación de procesos)

Qué datos mejorar

- Crear una lista ordenada por prioridad de datos a mejorar
  - Consideraciones posibles
    - Dar más importancia a la estrategia de negocio de la empresa
      - Ej.: empresa que está apuntando al marketing directo, debería priorizar datos de clientes.
      - Ej.: empresa que está enfocada a mejorar eficiencia de operaciones, debería priorizar datos logísticos
    - Asociación con problemas del negocio ya conocidos
      - Ej.: reuniones perdidas con clientes: direcciones incorrectas, etc.
    - Tasas de errores reales vs. requerimientos de nivel de calidad
    - Económicas
      - Hay errores que tienen consecuencias más costosas que otros.

Completitud

- Limpieza
  - Adquisición de nuevos datos
    - Tarea muy costosa
  - Imputación (Densidad)
    - Aplicación de reglas de consistencia
    - Mantener la distribución de la frecuencia de los valores en cada atributo
    - Mantener características del conjunto de datos (media, varianza, etc.)
    - Explotar dependencias funcionales
      - Ej.: nro_dormitorios → ingreso
    - Aplicar técnicas estadísticas y métodos de *machine learning*
      - Regresión lineal, SVM, etc (Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B., & Tabona, O. (2021). A survey on missing data in machine learning. *Journal of Big Data*, *8*(1), 1-37.)

Completitud

- Eliminación de Causas
  - Manejador de Bases de Datos (Densidad)
    - NOT NULL
  - Analizar causas para poder hacer cambios (Cobertura)
    - Ejemplos
      - No hay información de ventas durante 3/1 .. 3/4 ?
      - No hay productos con precio > 20 ?
      - Datos truncados y censurados
        - » Datos truncados y censurados
        - » Ventas de más de $100000 se guardan como $100000
    - Siempre se necesita conocimiento del dominio

Exactitud – Exactitud Sintáctica

- Limpieza
  - Normalización/Estandarización
    - Conversión de tipo de datos. Ej.: varchar → int
    - Normalizar: llevar a un formato común
      - date: 03/01/05 → 01-MAR-2005
      - moneda: $ → €
      - Mayúsculas / minúsculas
      - tokenizing:
        - » “Martínez, Cristina” → “Martínez”, “Cristina”
        - » direcciones: facilita comparaciones
    - Discretizar valores numéricos
    - Transformaciones específicas del dominio

Exactitud

- Limpieza
  - Uso de diccionarios y referenciales
  - Herramientas de limpieza de dominios específicos
    - Limpieza de calles, de nombres, de ciudades, etc.

- Eliminación de Causas
  - Manejador de Bases de Datos (Exactitud sintáctica)
    - Def de tipos de datos y restricciones de dominio
    - Restricciones tipo “Check”

Consistencia

- Limpieza
  - Imputación satisfaciendo reglas de consistencia
    - Cambiando la menor cantidad de valores posible y manteniendo la distribución de la frecuencia de los valores en cada campo.
    - Ejemplo
      - (Edad, EstadoCivil, TipodeTrabajo)
      - <68, casado, jubilado> → (error) <6, casado, jubilado>
      - Existe regla: Edad < 15 → EstadoCivil ≠ casado
      - Podemos corregir poniendo 15 en vez de 6, respetando el mínimo cambio (1er. objetivo), pero si lo hacemos muchas veces vamos a variar la frecuencia relativa.
    - Puede haber reglas implícitas que se derivan lógicamente de las explícitos

Consistencia

- Eliminación de Causas
  - Manejador de Bases de Datos
    - Integridad de dominio
      - Def de tipos de datos y restricciones de dominio
      - Restricciones tipo “Check”
      - Triggers
    - Integridad intra-relación
      - Primary Key, Unique
    - Integridad inter-relación
      - Foreign Key

Unicidad

- Limpieza
  - Se hace cuando hay conflictos de datos en registros duplicados
  - Técnicas de Fusión de datos (integración de datos)
    - Elegir datos más confiables en base a su procedencia
    - Mantener todos los datos (por ej., todas las direcciones que se encontraron para un mismo cliente)
    - Llenar datos faltantes en un registro con datos existentes en el duplicado
- Eliminación de Causas
  - Manejador de Bases de Datos
    - Unique, Primary Key

Frescura

- Limpieza
  - Actualización de datos
- Eliminación de Causas
  - Cambiar frecuencia de actualización de los datos
  - Manejador de Bases de Datos
    - Replicación, Vistas Materializadas

Limpieza

- Ordenar las tareas de limpieza para mejorar los resultados
  - Un orden razonable podría ser
    - Frescura
    - Completitud
    - Exactitud
    - Consistencia
    - Unicidad
- Definir dónde y cómo se incluyen los datos que fueron corregidos en los procesos de la organización
  - dónde se almacenarán los datos limpios
  - cómo se disponibilizarán a las distintas aplicaciones

Herramientas de Limpieza

- Potter's wheel
  - Estandarización, profiling, limpieza para SID (sistemas de integración de datos)
- Telcordia's tool
  - Estandarización, limpieza para SID
  - Dominio: direcciones, impuestos
- Ajax
  - Normalización, limpieza para SID
  - Dominio: referencias bibliográficas
- Arktos
  - Estandarización, localización de errores, limpieza para SID
  - Dominio: ETL, aplicaciones de salud
- Choice Maker
  - limpieza para SID
  - Dominio: nombres, direcciones, negocios, datos médicos, datos financieros
- Intelliclean
  - Normalización, limpieza para SID

Referencias en el texto del curso (8.2)

Herramientas de Limpieza

Empresa	Productos
Ataccama	DQ Analyzer, Data Quality Center, DQ Issue Tracker, DQ Dashboard
Datactics	Data Quality Platform, Data Quality Manager, Master Record Manager
DataMentors	DataFuse, ValiData, NetEffect
Human Inference	HIquality Suite, HIquality Name Worldwide, HIquality Identify, HIquality Data Improver, DataCleaner
IBM	InfoSphere Information Analyzer, InfoSphere QualityStage, InfoSphere Discovery
Informatica	Data Explorer, Data Quality, Identity Resolution, AddressDoctor
Information Builders/iWay	iWay Data Quality Center
Innovative Systems	i/Lytics Data Quality, i/Lytics Data Profiling, i/Lytics ProfilerPlus, FinScan
Oracle	Oracle Enterprise Dat Quality, Oracle Enterprise Data Quality for Product Data
Pitney Bowes Software	Spectrum Technology Platform
RedPoint (DataLever)	RedPoint Data Management
SAP	Data Quality Management, Information Steward, Data Services
SAS/DataFlux	Data Management Platform
Talend	Talend Open Studio for Data Quality, Talend Enterprise Data Quality
Trillium Software	Trillium Software System, TS Discovery, TS Insight, Trillium Software On-Demand
Uniserv	Data Quality (DQ) Explorer, DQ Batch Suite, DQ Real-Time Suite, DQ Real-Time Services, DQ Monitor
Melissa Data	Contact Zone
Datiris	Datiris Profiler
CloverETL	Address Doctor
Microsoft	Data Quality Services

Manejador BDatos para evitar errores

SE EVITA	A TRAVES DE…
Tipos de datos incorrectos	Def de tipos de datos y restricciones de dominio
Valores erróneos	Restricciones tipo “Check”
Valores faltantes	“Not null”
Referencias inválidas	“Foreign Key”
Duplicados	“Unique”, “Primary Key”
Inconsistencias	Manejo de transacciones
Datos desactualizados	Replicación, Vistas Materializadas

Reflexión final…

- Para mejorar la calidad se debería aplicar
  - Prevención a través de correción de procesos para datos con alta frecuencia de creación y actualización (Eliminación de causas).
  - Corrección de errores para datos con baja frecuencia de creación y actualización (Limpieza).
  - Cualquier diseño de proceso o reingeniería debería luchar para que el nuevo proceso sea lo más libre de errores posible.