CALIDAD EN SISTEMAS DE INFORMACIÓN

Talend Open Studio

Práctica de Calidad de Datos

Anxo Pérez
anxo.pvila@udc.es

Information Retrieval Lab
Computer Science Department
University of A Coruña

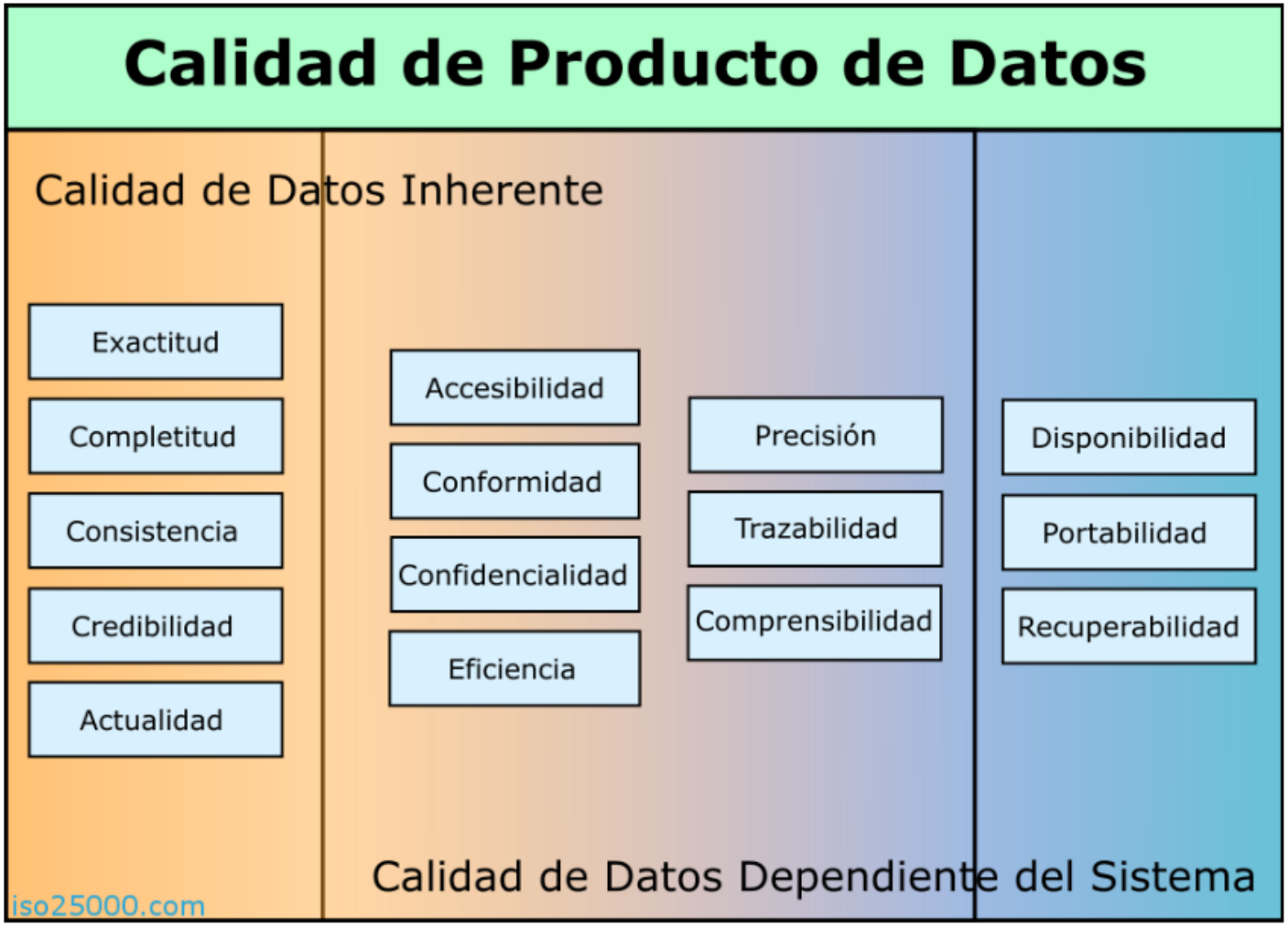
Práctica de Calidad de Datos: Resumen

- **Contexto:** La dirección está preocupada por la calidad de los datos corporativos (bases de datos internas). Se aproxima una auditoría, y hace mucho que delegamos a los administrativos la introducción de datos. Nos encargan evaluar y documentar la **calidad real** de dichos datos.
- **Objetivo general:**
  - Realizar un **análisis profundo** sobre las dimensiones de calidad (ISO 25012) en una base de datos proporcionada (MySQL).
  - Redactar un **informe** que describa los hallazgos y proponga **mejoras** concretas.
- **Herramienta principal: Talend Open Studio for Data Quality** para *data profiling, detección de errores, limpieza* y obtención de métricas.

Práctica de Calidad de Datos: Objetivos específicos

- **Estudio de la calidad de datos** según ISO 25012:
  - Revisar dimensiones clave (exactitud, completitud, consistencia, etc.).
  - Explicar cómo la falta de calidad puede afectar a procesos de negocio y decisiones.
- **Informe detallado**:
  - Identificar principales deficiencias (ej. valores nulos, duplicados).
  - Explicar impacto en la organización (costes, riesgos).
  - Proponer **recomendaciones** y planes de acción.
- **Formato de entrega**:
  - Informe en **PDF**, incluyendo nombres de **todos los miembros** (grupos de 3).
  - Solo un integrante sube el documento final al campus virtual.
  - Fecha de entrega: se pondrá la tarea en el campus virtual.

Práctica de Calidad de Datos: Objetivos específicos



¿Por qué Talend Open Studio for Data Quality?

- **Herramienta específica de Data Quality:**
  - Ofrece *data profiling*, detección de duplicados, creación de **reglas** para validar o limpiar datos.
  - Soporta **múltiples conectores** (archivos, DBs relacionales, etc.).
- **Ventajas**:
  - **Interfaz gráfica** que facilita la curva de aprendizaje.
  - Permite explorar **métricas** (porcentaje de nulos, valores fuera de rango, análisis de patrones).
- **Usos en la industria**:
  - Muchas empresas utilizan Talend (o herramientas similares) para proyectos de **ETL** y control de calidad.
  - Principal objetivo: Familiarizarse con este tipo de software.

Pasos con Talend Open Studio

1. Instalación y tutoriales

- Descargar **Talend Open Studio for Data Quality** (**Link**).
  - Revisar la documentación oficial: **Documentación Talend**.
  - Revisar la guía de usuario subida al campus virtual.

2. Preparación de la base de datos

- Descargar la base de datos subida al campus virtual (`.sql`).
- Importar la bases de datos en un servidor de MySQL local:

```
mysql -u root -p tbi_new_db <
"tbi_new.sql"
```

3. Conexión en Talend

- Configurar la conexión a MySQL en Talend. Aquí es necesario añadir en el campo de Additional parameters el parámetro &serverTimezone=UTC.

Desarrollo del análisis en Talend

- **Data profiling**:
  - Calcular **estadísticas** (distribuciones, cardinalidad, % n distinct valores vacíos, etc.).
  - Identificar **columnas críticas** con alta proporción de NULLs o valores fuera de rango.
- **Reglas de calidad**:
  - Crear **indicadores** que representen constraints (p.ej. **price > 0** o **id `not null`**).
  - Detectar **incumplimientos** y **cuantificarlos**.
- **Detección de duplicados / Fuzzy matching** (opcional):
  - Si hay columnas de **identificación** (nombres, emails, etc.), ver si existen duplicados (record linkage).