Introducción al concepto de los datos y su análisis

Tema 1

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Índice

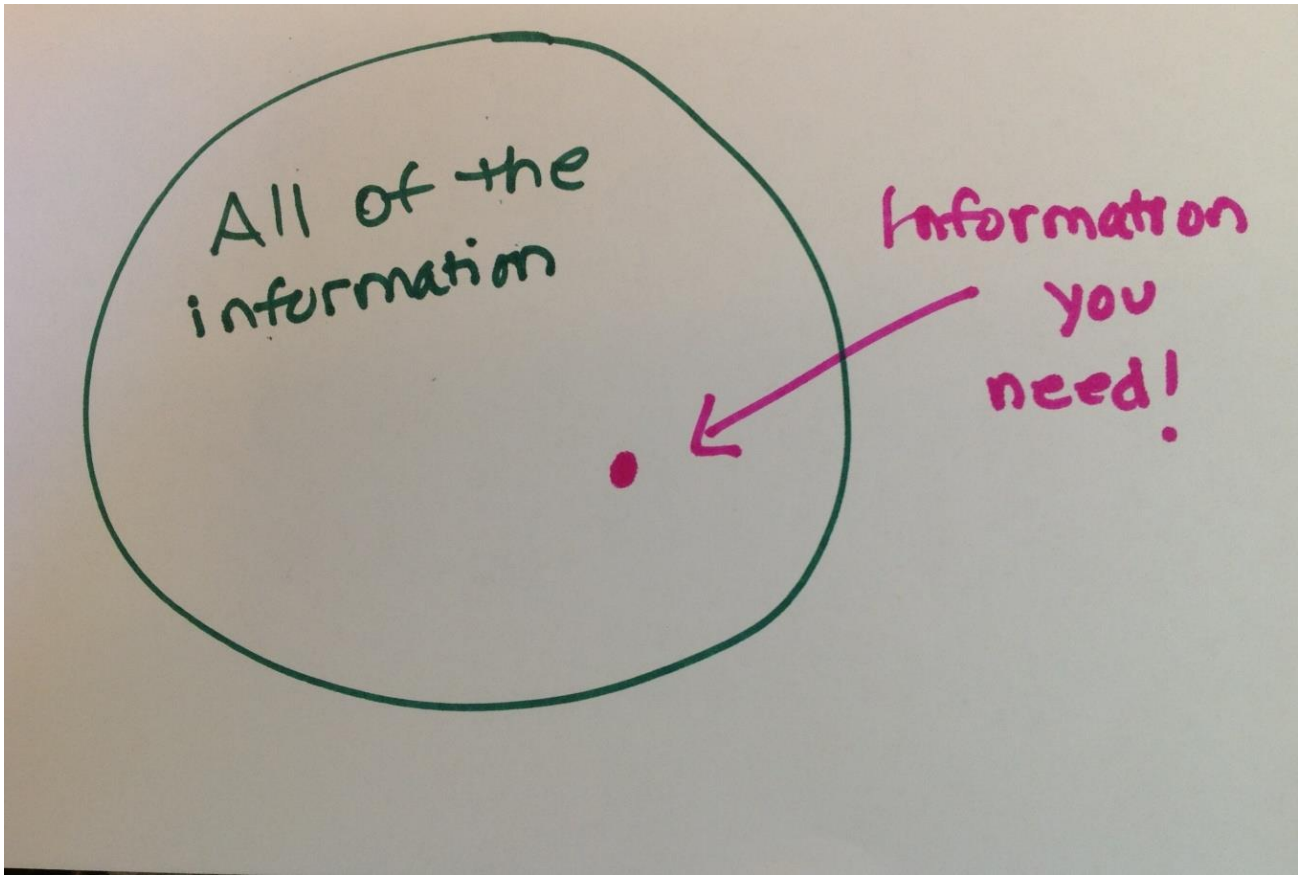
- Planificación de la asignatura en Moodle de la UA
  - Acceso a través de UACLOUD

Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

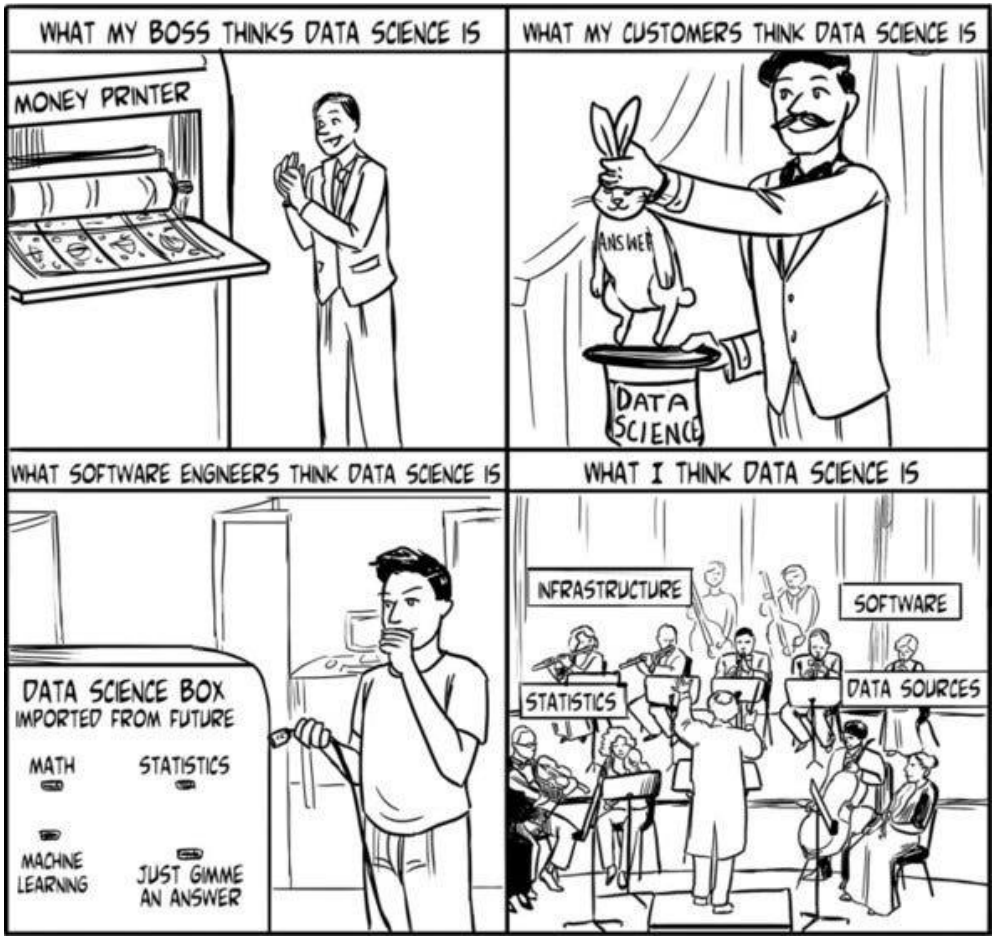
Introducción a Big Data

- Introducción a Big Data



Introducción a Big Data

- ¿Qué es la Ciencia de datos?



Introducción a Big Data

- Un proyecto de Big data no es un proyecto de una única persona
  - Servidores
  - Arquitectura
  - Programación
  - Diseño
  - Análisis
  - Dirección
    - DevOps, Backend, Frontend, Data scientist…

Introducción a Big Data

- ¿Quién realiza esta tarea?
  - Ingeniero de datos
    - Lidia con problemas de recolección, gestión, transformación y publicación de datos
    - Tiene conocimientos de: Modelado, SQL, NoSQL, ETL/ELT, Spark, MapReduce, Kafka, Flume,
    - Desarrollo y despliegue en la nube…

Introducción a Big Data

- ¿Quién realiza esta tarea?
  - Data scientist
    - …a data scientist is 1) a data analyst in California or 2) a statistician under 35 (Gartner blog post by analyst Svetlana Sicular).
    - Lidia con problemas de recolección, gestión, transformación y publicación de datos
    - Tiene conocimientos de: Modelado, SQL, NoSQL, ETL/ELT, Spark, MapReduce, Kafka, Flume, Estadística R, Matlab, SAS, SPSS, Minería de datos, Procesamiento de lenguaje natural, Machine Learning, Map/Reduce, Hadoop, Hive, Python, Desarrollo y despliegue en la nube…
  - La separación entre ambas "figuras" no siempre es clara a la hora de abordar un proyecto

Introducción a Big Data

- Introducción a Big Data
  - The notion of a Data Scientist is a little mad but then so is Big Data.

Removing the buzzwords just leaves you with….Data.

Introducción a Big Data

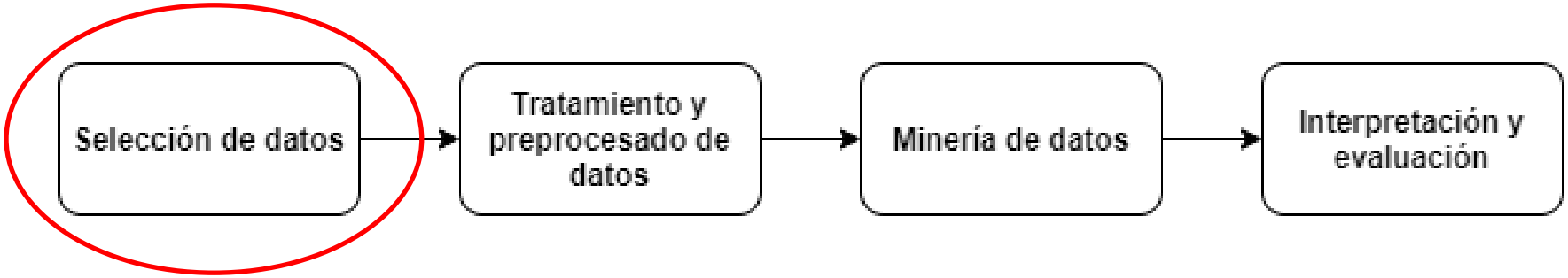
- Introducción a Big Data
  - Con el auge del "Big Data" y el crecimiento de técnicas de inteligencia artificial, ha hecho que el ML sea una practica común tanto en educación como en empresa.
  - Un Proyecto de Machine Learning (ML) requiere no solo conocimientos en técnicas de procesamiento de datos y algoritmos, si no también conocimiento profundo del dominio.
  - Fuerte dependencia y necesidad de comunicación entre desarrolladores y stakeholders donde además se deben de capturar numerosos requerimientos (funcionales y no funcionales)

Introducción a Big Data

- Big Data e Inteligencia artificial
  - Uno de los principales problemas para el éxito de los proyectos analiticos, es que los proyectos de data warehouses, visualizaciones de usuario y proyectos de ML han sido abordados de manera independiente.
    - Ignorando posibles requerimientos cruzados
    - Restricciones colectivas
    - Dependencias entre salidas de diferentes sistemas
    - Objetivos de negocio

Introducción a Big Data

- Ciclo de vida del dato



Introducción a Big Data

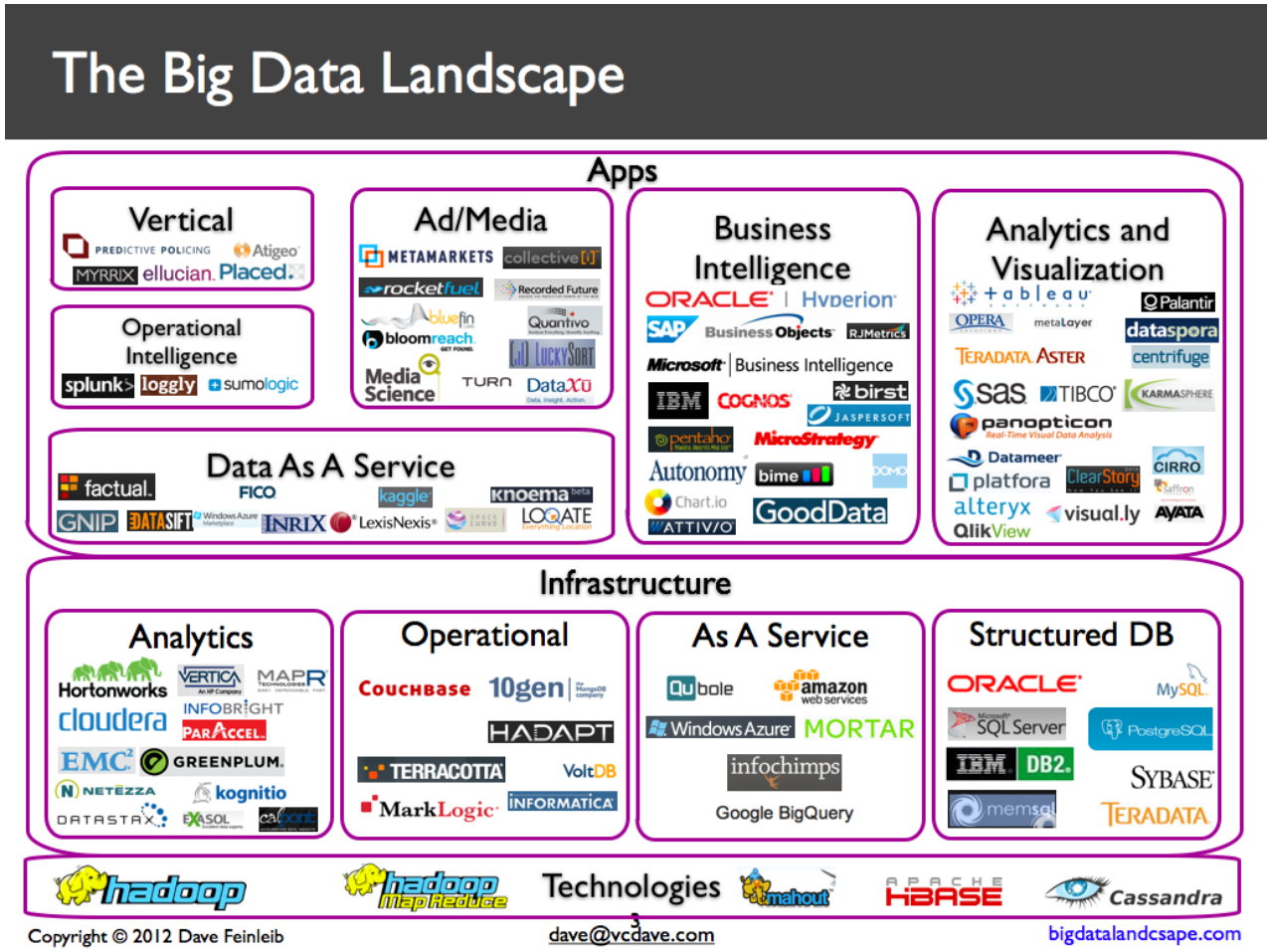
- Adquisición de datos
  - Tratamos de obtener tantos datos de valor como sea posible, esto incluye fuentes internas de la organización, como fuentes externas.
    - Bases de datos, ficheros Excel, informes, APIs…
  - El objetivo es obtener los datos necesarios para el análisis, automatizando en la medida de lo posible la extracción de los mismos (ETLS).
  - En este punto "no importa" tanto los datos si no las entidades y fuentes.

Introducción a Big Data

- BIG DATA para salvar el mundo
  - Siempre hemos tenido mucha información
  - Pero ahora gracias a nuevas herramientas y técnicas se pueden analizar e interpretar
  - También se pueden almacenar más cantidad de información
    - Genoma Humano
    - Datos de Enfermedades
    - LHC

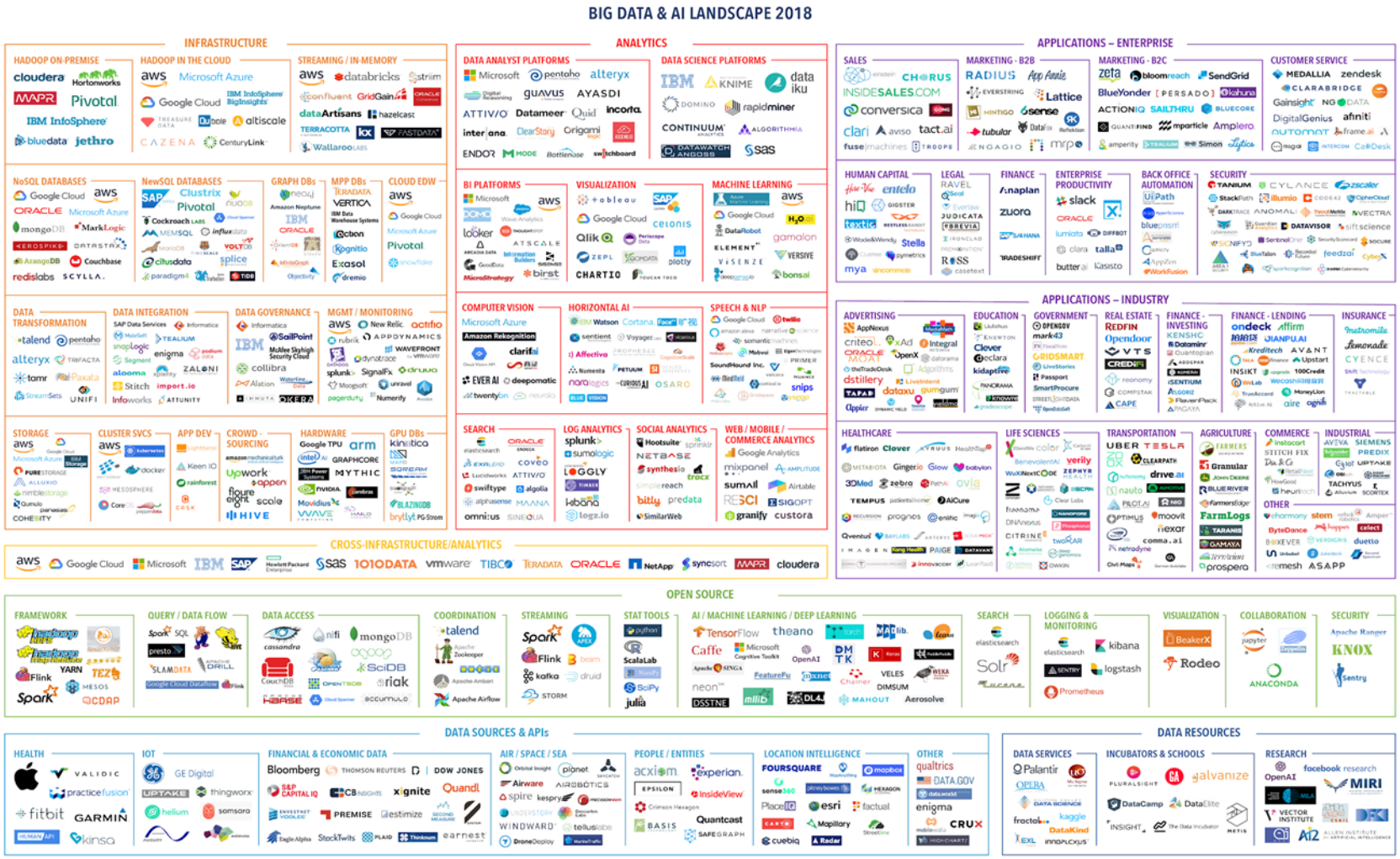
Introducción a Big Data

- Ya hay muchos jugadores…



Introducción Big Data

2018



Final 2018 version, updated 07/15/2018

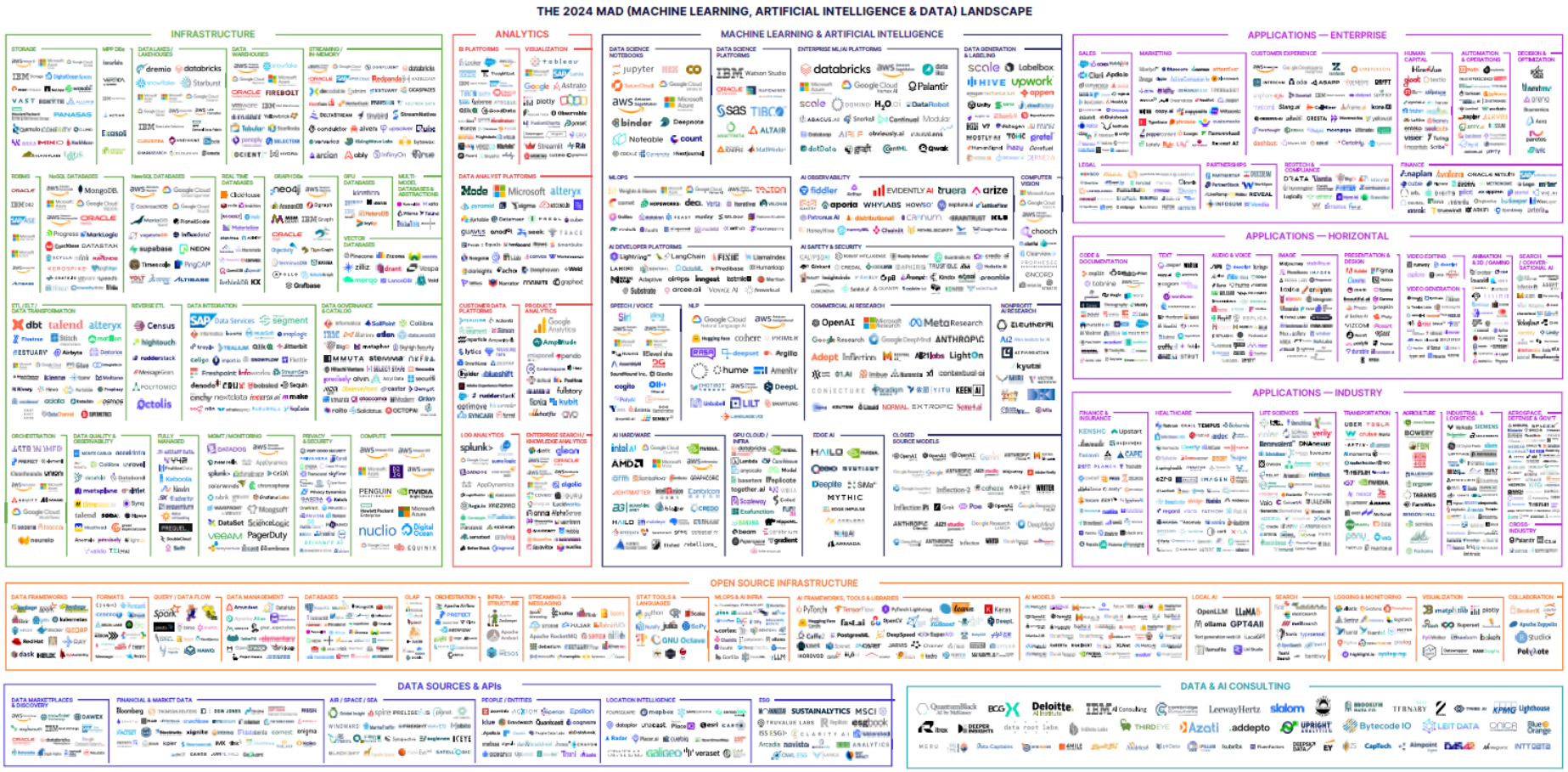
© Matt Turck (@mattturck), Demi Obayomi (@demi_obayomi), & FirstMark (@firstmarkcap)

mattturck.com/bigdata2018

FIRSTMARK EARLY STAGE VENTURE CAPITAL

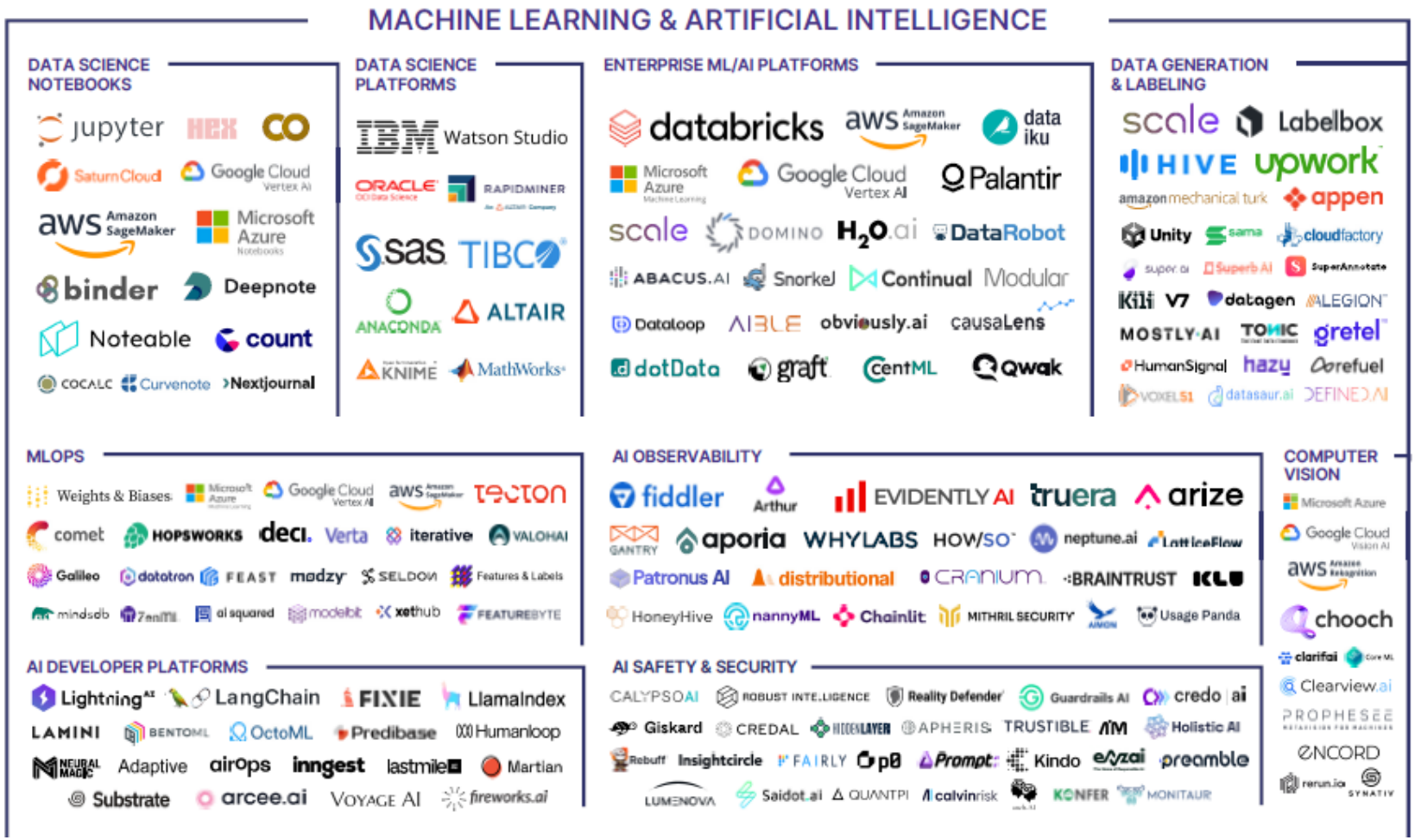
Introducción Big Data

2024



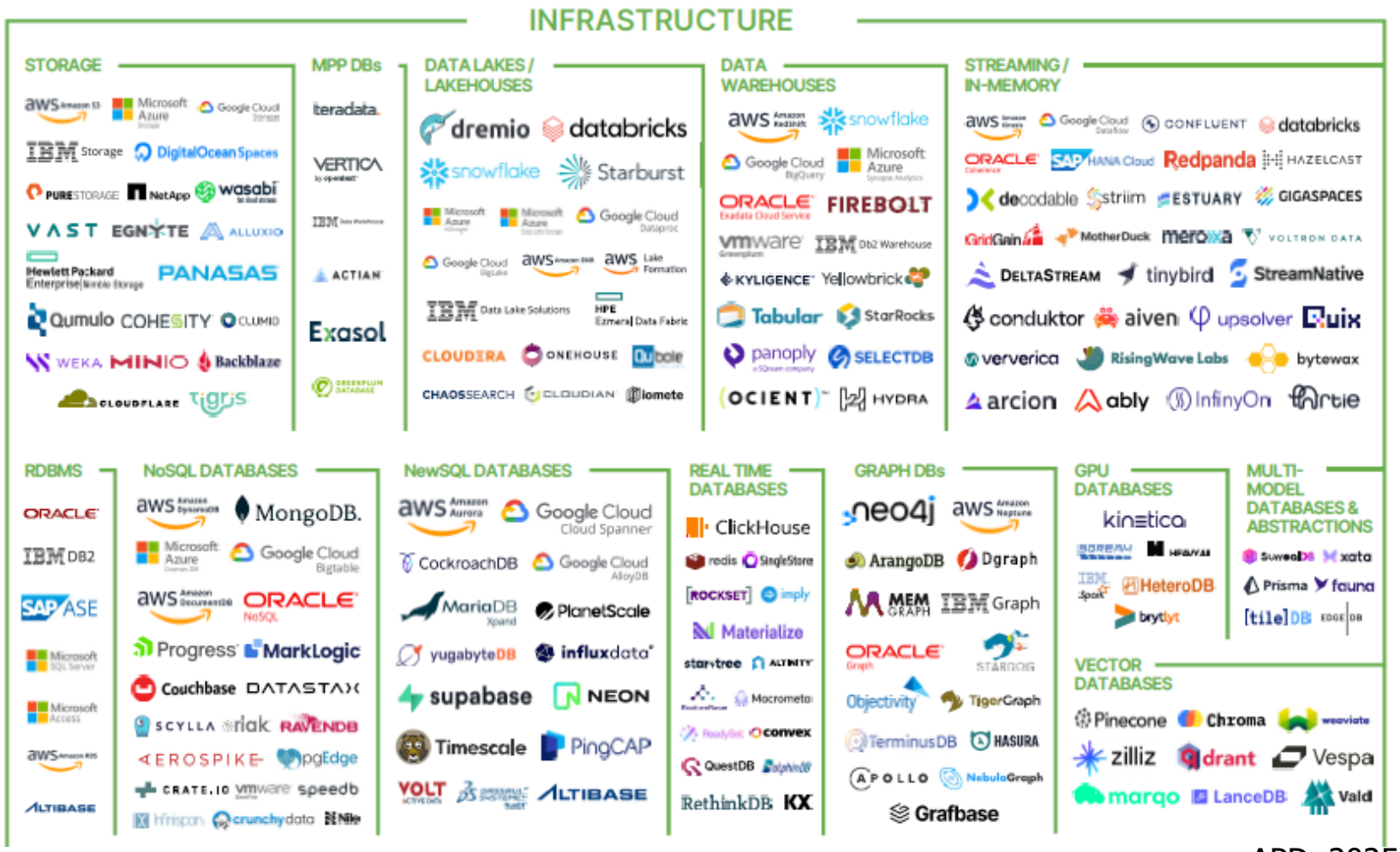
Introducción Big Data

2024



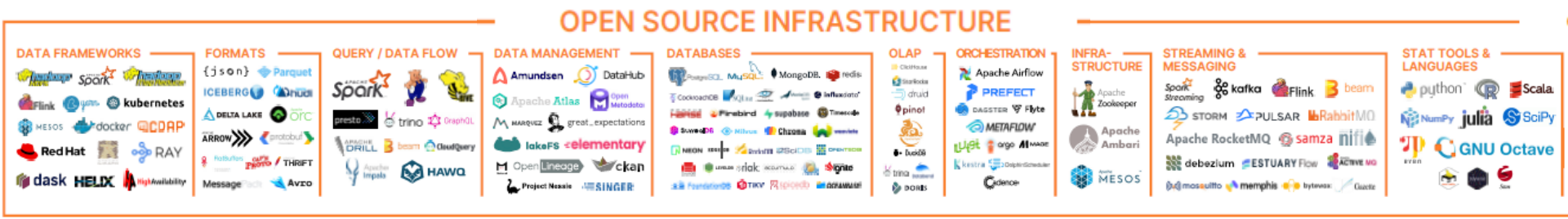
Introducción Big Data

2024



Introducción Big Data

2024



Introducción a Big Data

- Dentro de Big Data se engloba o tiene que ver:
  - Smart City
  - Sensores
  - Seguridad
  - Privacidad
  - Inteligencia artificial
    - …
  - Tendencias
  - Marketing
  - Psicología
    - …

Introducción a Big Data

- ¿Qué es Big data?

  "Forma de afrontar el procesamiento o análisis de grandes volúmenes de información que por su naturaleza desestructurada no pueden ser analizados, y en un tiempo aceptable, usando los procesos y herramientas tradicionales de BI" (IBM)

- Características (5v's)

Veracidad

Velocidad

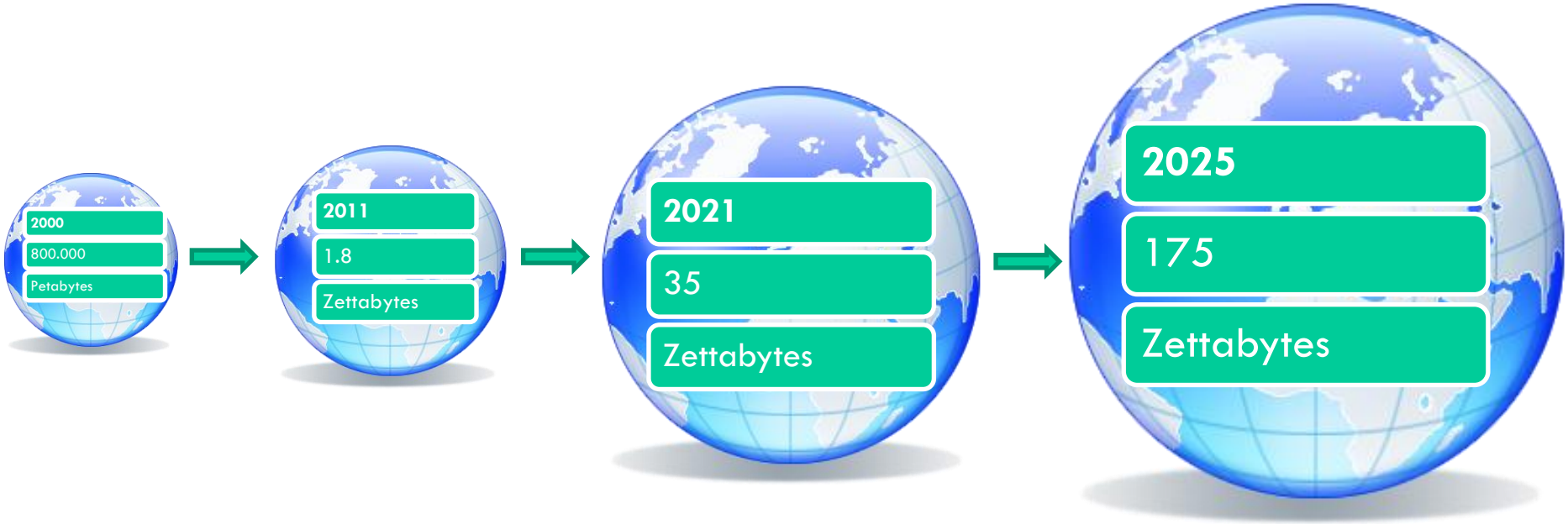
Valor

Volumen

Variedad

Introducción a Big Data

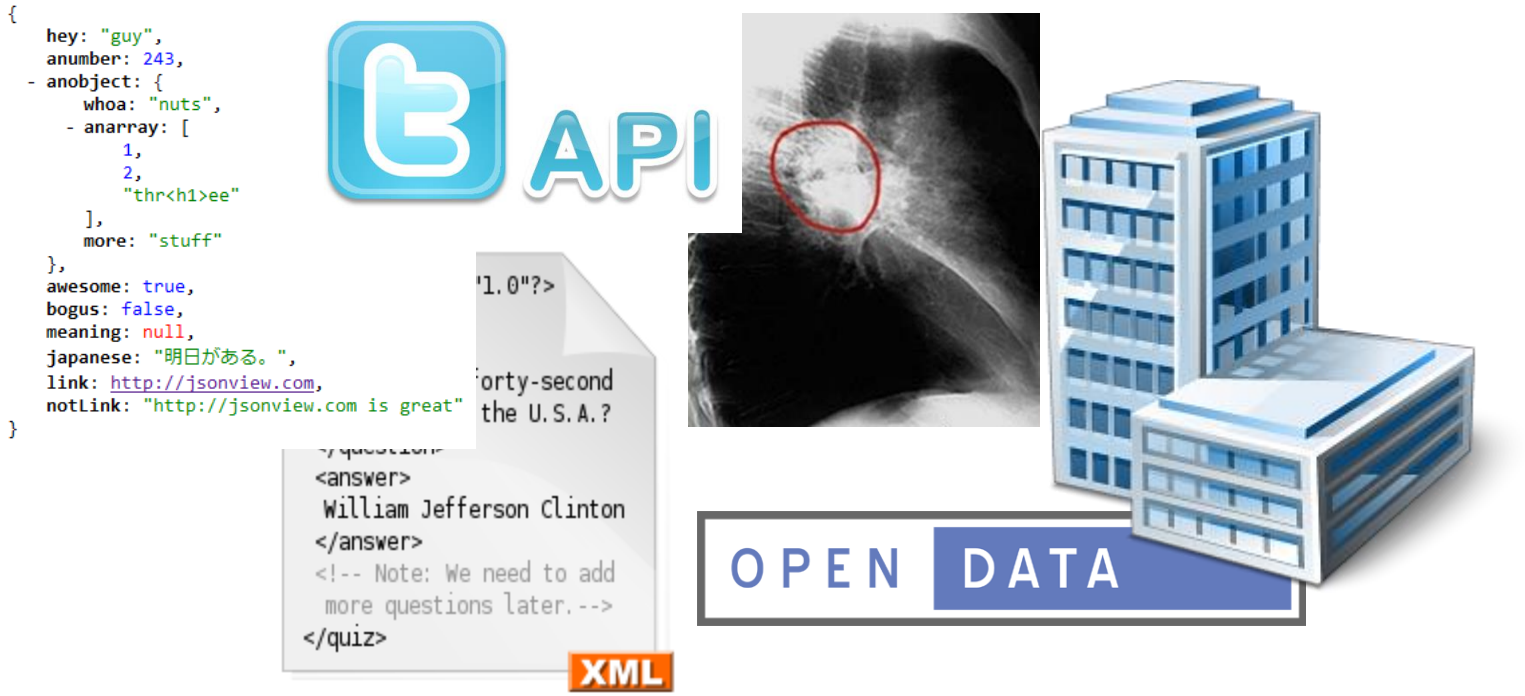
- Volumen: capacidad para procesar **grandes volúmenes de datos**



1 Zettabyte = 10^6 Pettabytes = 10^9 Terabytes = 10^{12} Gigabytes

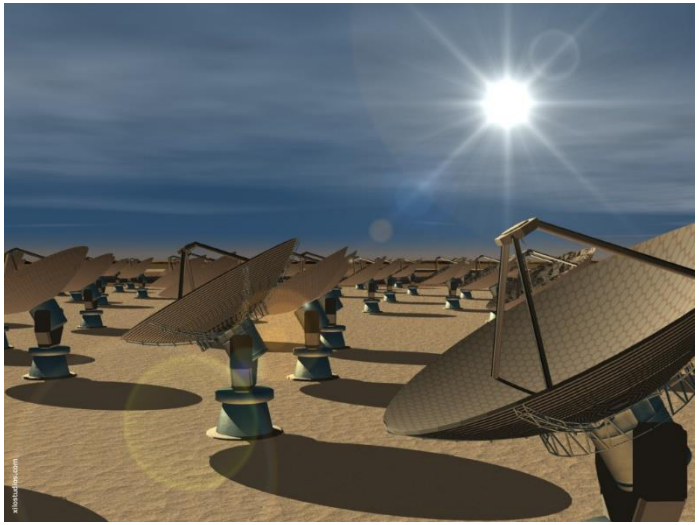
Introducción a Big Data

- Variedad: capacidad para soportar el aumento en la **heterogeneidad** de las **fuentes** a procesar.



Introducción a Big Data

- Velocidad: velocidad a la que fluye la información.

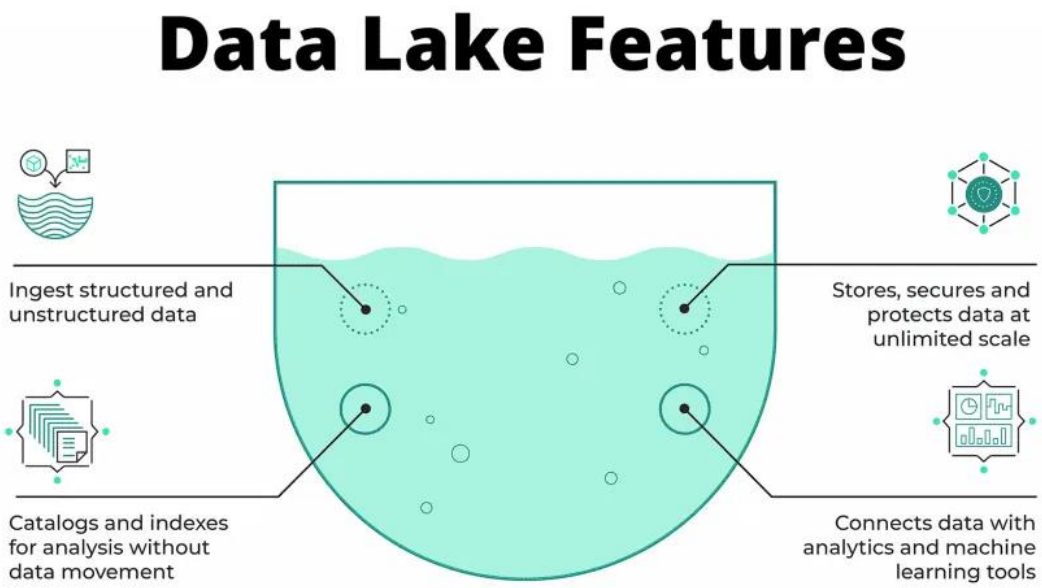


Telescopio SKA
10 petabytes / hora

X
350.000 por min
200 billones por año

Introducción a Big Data

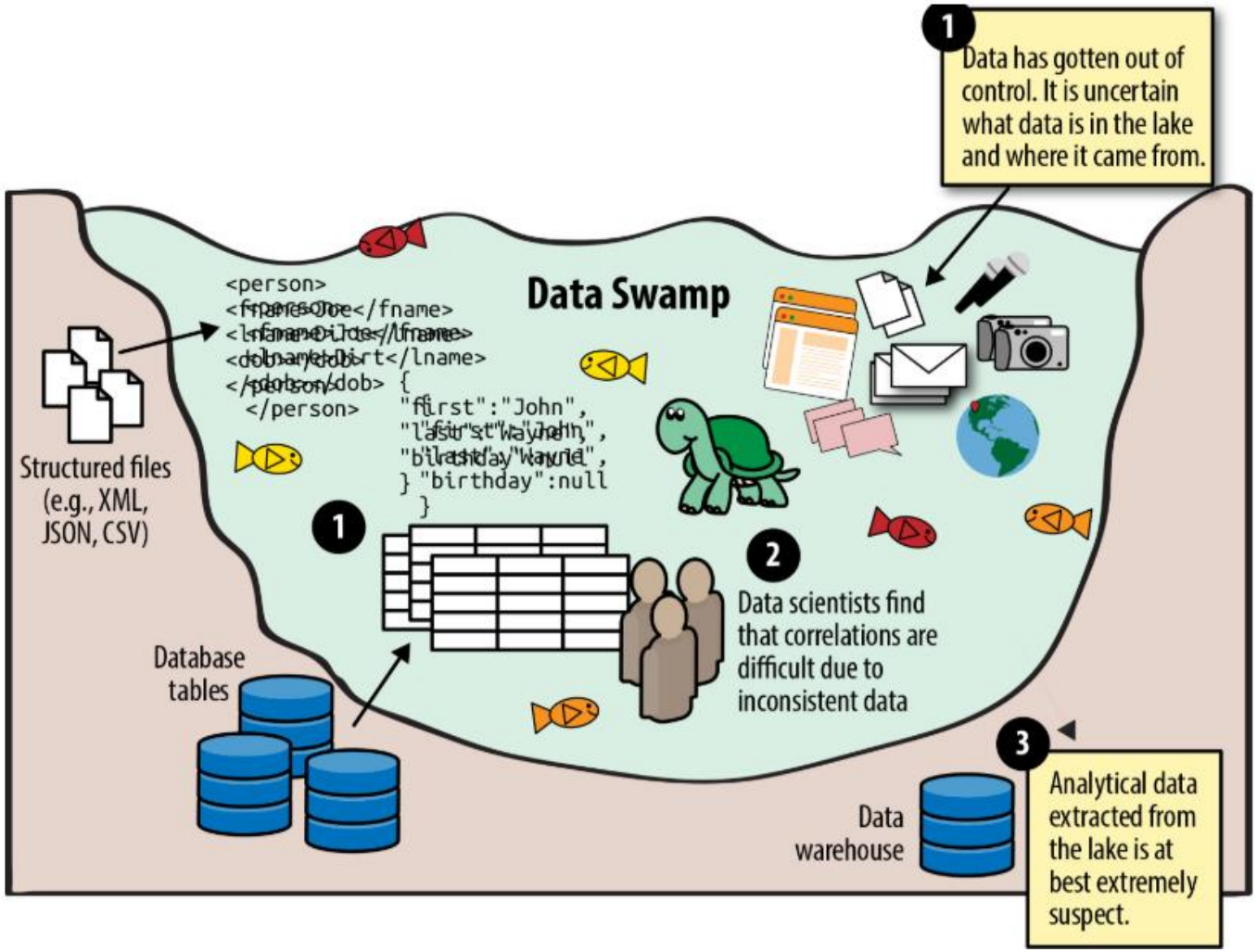
- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Lake: Almacén centralizado de datos en bruto (raw data)



Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Lake: Almacén centralizado de datos en bruto (raw data)
    - Múltiples formatos: estructurados, semiestructurados y no estructurados
    - Escalable y flexible, pensado para grandes volúmenes de datos
    - Ideal para análisis exploratorio, ciencia de datos e IA
    - Bajo coste de almacenamiento, pero ¡requiere buena gobernanza y gestión o se convertirá un data swamp!

Introducción a Big Data

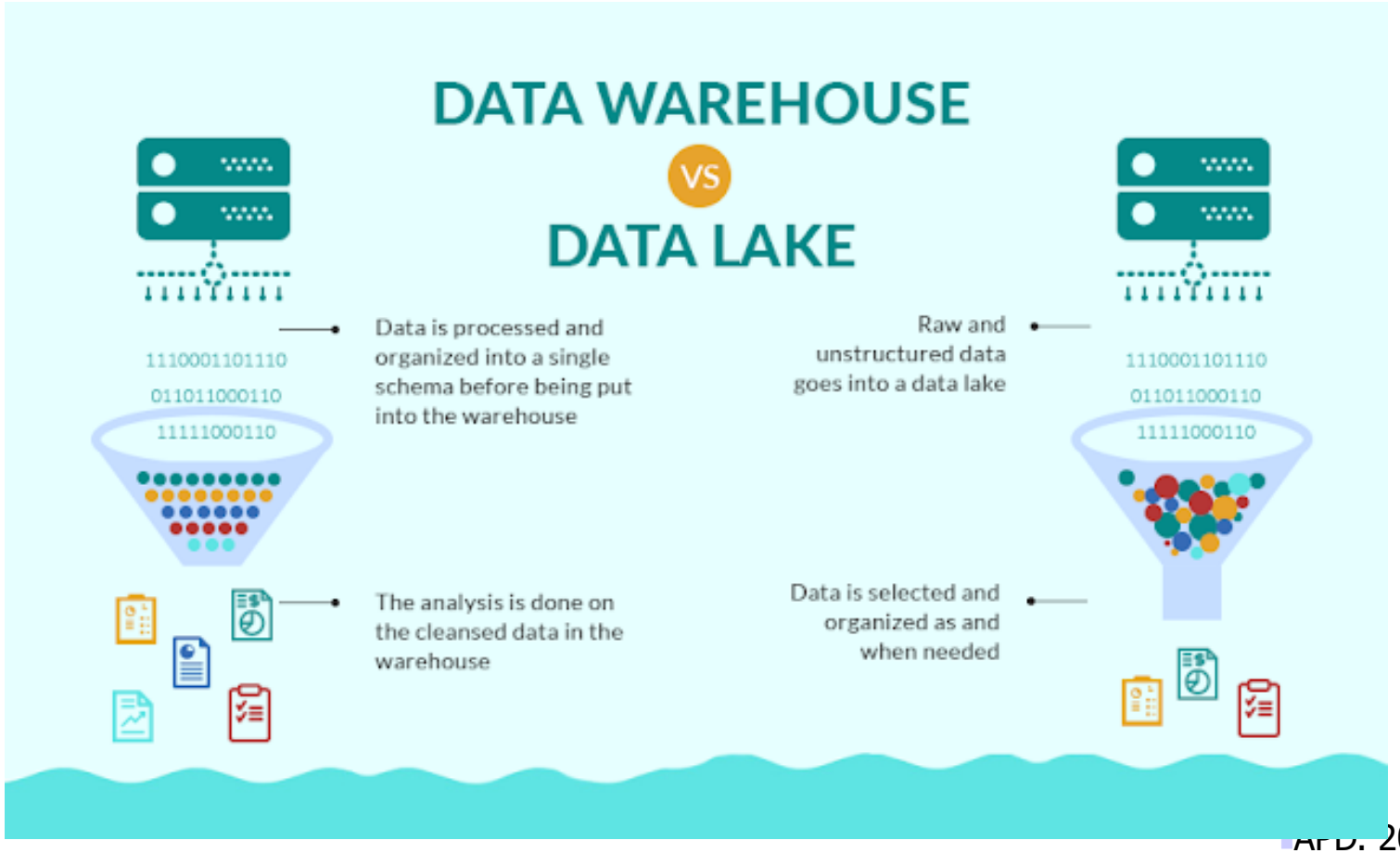


Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?
  - Data Swamp: ¡Es un Data Lake MAL GESTIONADO!
    - Datos desorganizados, falta de metadatos ni controles de calidad.
    - Difíciles de explorar (¿Qué son estos datos?), de mantener (de donde vienen estos datos?, y de extraer valor (¿Para qué nos sirve esto?)
    - Duplicidades, inconsistencias -> Perdida de confianza

Introducción a Big Data

- ¿Dónde se guarda toda esa ingente cantidad de datos?



Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

Nuevas necesidades de gestión de datos

- Situación actual: Entornos económicos altamente competitivos
  - Típica pregunta a SGBDR
    - ¿Cuántos zapatos vendimos el último mes?
  - Empresas necesitan adoptar decisiones estratégicas
    - ¿Cuántos zapatos del 41 de color rojo se vendieron el último mes en la zona norte, este y sur; comparados con las ventas del mismo mes el año pasado ?
    - ¿Qué tipo de cliente me ha estado comprando el BMW 320i durante los últimos 10 años ?
      - ¿Directivo, profesor, trabajador escala básica,...?
  - Empresas quieren conocer que pasará
    - ¿Cuántos zapatos venderé…? ¿Quién comprará…?

Nuevas necesidades de gestión de datos

- Req. 1. Gran volumen de datos (varios años, clientes, productos, almacenes, etc.)
  - Históricos y procedentes de distintas fuentes

- Req. 2. Tienen que ser presentados en un entorno amigable y fácil de usar
  - Entender el tipo de preguntas "estratégicas"

Nuevas necesidades de gestión de datos

- ¿Son válidos los sistemas OLTP para tales decisiones?
  - Problemas
    - Datos históricos no disponibles en sistemas diarios OLTP (OnLine Transaction Processing)
    - Normalmente en distintas fuentes de datos
      - Proveedores, Clientes, componentes, productos defectuosos, etc.
    - Los directivos no saben manejar tales sistemas y,
    - Rendimiento, errores, etc….

Nuevas necesidades de gestión de datos

- El uso de OLTP requeriría:
  - Integrar datos → Consumir tiempo (req. 1)
  - ¿Podría un analista manejar herramientas transaccionales ? (req. 2)

- Vamos a hacer breve historia sobre las soluciones planteadas para manejar datos históricos

Nuevas necesidades de gestión de datos

- Datos procedentes de sistemas heredados
  - 1970's se hace uso de IBM mainframes
    - Cobol, CICS, IMS, DB2, etc
  - 1980's plataformas AS/400 y VAX/VMS
  - Hoy en día muchas aplicaciones de negocio "corren" sobre estos sistemas
    - Muchos años recogiendo datos y reglas de negocio → dificultad para llevarlos a otro sistema
    - Los datos se vuelcan en librerias donde accederán otras aplicaciones de negocio
  - Coste de aplicaciones de negocio es grande

Nuevas necesidades de gestión de datos

- Datos extraídos en el escritorio (1990's)
  - Se reduce la distancia entre usuario final y programador
    - PC con hojas de cálculo, herramientas de análisis, etc.
  - Herramientas de análisis que acceden a datos producidos por sistemas heredados
  - Problema: datos permanecen fragmentados y están orientados a necesidades específicas de usuarios final
    - Soluciones parciales
    - No todos usuarios tienen la misma destreza

Nuevas necesidades de gestión de datos

El Excel se ha usado en la Fórmula 1 hasta que se han dado cuenta que no es la mejor forma de controlar las 20.000 piezas del coche

James Vowles, nuevo jefe de Williams, encontró uno de los motivos por los que el histórico equipo de la Fórmula 1 estaba tan atrasado

2025

Nuevas necesidades de gestión de datos

- Sistemas de apoyo a la decisión (SAD) y Sistemas de Información de Ejecutivos (SIE):
  - SAD: Datos en detalle. Ejecutivos medios y bajos
  - EIS: Datos consolidados. Altos ejecutivos
    - Más orientados a vista multidimensional de los datos
  - Son similares y solapan funcionalidades
  - Son los precursores de los Almacenes de datos
  - Alto precio y de nuevo descoordinación sobre los datos necesarios para efectuar el análisis

Nuevas necesidades de gestión de datos

- Resumen características SAD y EIS:
  - Datos descritos en términos de negocio, en lugar de términos técnicos como tupla, fila o tabla.
    - Sistemas enfocados a usuarios no técnicos
  - Datos preprocesados siguiendo patrones de reglas de negocio
    - Beneficios por la venta de productos en distintos almacenes
  - Vistas consolidadas de datos
    - Aunque permiten ver datos en detalles, rara vez pueden acceder a todos los datos en detalle

Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

El almacén de datos: primera aproximación

- Evolución de los sistemas de almacenes de datos
  - Desde los primeros sistemas de gestión de ficheros (a. 1970) hasta los Sistemas de Gestion de Bases de Datos actuales (SGBD)
    - SGBD eficientes
    - SGBD robustos
    - Múltiples herramientas de alto nivel que facilitan su manejo
      - Servidor
      - Cliente

El almacén de datos: primera aproximación

- Empresas en la actualidad
  - Aproximadamente el 90% de SGBD son relacionales
    - Integridad de los datos
    - Independencia de los datos, etc.
  - Orientados a los procesos diarios de la empresa
    - Sistemas de Procesamiento Transaccional en Línea (*On-Line Transactional Processing, OLTP*)
    - Compras de productos, ventas, pedidos, gestión de clientes, etc.
  - Datos históricos → almacenamientos externos (Data Centres, almacenamiento en la nube, discos externos,…)

El almacén de datos: primera aproximación

- El Almacén de datos (Data Warehouse, DW)
  - Sistemas que almacenan datos históricos para ser utilizados por los SAD
  - Son sistemas eminéntemente de consulta enfocados a extraer conocimiento de los datos históricos almacenados
  - El análisis de los datos → On-Line Analytical Processing (OLAP)
    - Utilizan el modelado multidimensional (cubos, hipercubos, etc)

El almacén de datos: primera aproximación

- Definición según W. Inmon (1992)
  - "Una colección de datos orientados por tema, variables en el tiempo y no volátiles que se emplea como apoyo a la toma de decisiones estratégicas"

El almacén de datos: primera aproximación

- Orientados por tema
  - El diseño enfocado a responder eficientementa a consultas estratégicas
    - Actividades de interés: compra, ventas, alquileres,…
    - Contexto de análisis: clientes, vendedores, productos, etc…
  - El modelado Multidimensional (primera aproximación)
    - Hechos → actividades de interés
    - Dimensiones → contexto de análisis

El almacén de datos: primera aproximación

- Integrados
  - Datos integrados de distintas fuentes de datos operacionales
- Variables en el tiempo
  - Datos relativos a un periodo de tiempo y se incrementan periódicamente
- No volátiles
  - Los datos almacenados no se modifican ni actualizan nunca, sólo se añaden nuevos datos

El almacén de datos: primera aproximación

- Objetivo técnicas de almacenamiento de datos (Data Warehousing)
  - Ofrecer información histórica para la toma de decisiones integrando la información procedente de distintas fuentes de datos operacionales

El almacén de datos: primera aproximación

- Las tecnologías de almacenamiento de datos integran eficientemente tecnología de bases de datos con análisis de datos
  - Bases de Datos: Un SGBD que soporte el almacén
  - Análisis de datos: herramientas que permitan llevar a cabo un análisis sencillo de los datos
    - Las más extendidas: herramientas OLAP
    - Análisis Multidimensional basado en el modelo multidimensional

El almacén de datos: primera aproximación

- Ejemplo: conocer las ventas de productos con respecto a los productos vendidos, los almacenes donde se vendieron y el tiempo

Ventas			Producto.Grupo = "Supermercado"			
			Comida		Bebida	
			Cong	Fresco	Refresco	Alcohol
Almacén. comunidad = "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Sagunto	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

El almacén de datos: primera aproximación

- Ventajas para las empresas
  - Decisiones soportadas por datos fiables
  - Rentabilidad de inversiones
  - Aumentar la competitividad en los nuevos entornos hostiles
  - Entornos amigables → los directivos analizan por ellos mismos los datos
    - Por fin se entienden con los ordenadores → lo hemos conseguido !!!

El almacén de datos: primera aproximación

- Inconvenientes
  - Infravalorar los recursos necesarios para alimentar el DW a partir de datos operacionales
  - No planificar el alto esfuerzo para lograr un buen diseño
  - Nunca está acabado → incremento continuo de requerimientos ad-hoc
  - ¡¡¡¡ Cuidado !!!! A más cantidad de datos → Más seguridad

Índice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

Sistemas operacionales VS Analíticos

	OLTP	OLAP
Usuario	Profesional de TI	Analista de Información
Función	Operaciones diarias	Apoyo a la decisión
Diseño de BD	Orientada a aplicación (Basado en EE-R)	Orientado al tema (esquema estrella, copos)
Datos	Actuales, Aislados	Históricos, Consolidados
Vistas	Detallados, Planos, Relac.	Agregados
Destino/utilización	Estructuradas, repetitivas	Ad-Hoc
Unidades de trabajo	Transacciones simples	Consultas complejas
Acceso	Lectura/escritura	Lectura principalmente
# Registros accedidos	Decenas	Millones
# Usuarios	"Miles"	"Centenares"
Tamaño de la BD	100 MB-GB	100 GB-TB
Medidas de rendimiento	Nº transacciones	Nº consultas, Respuesta

Indice

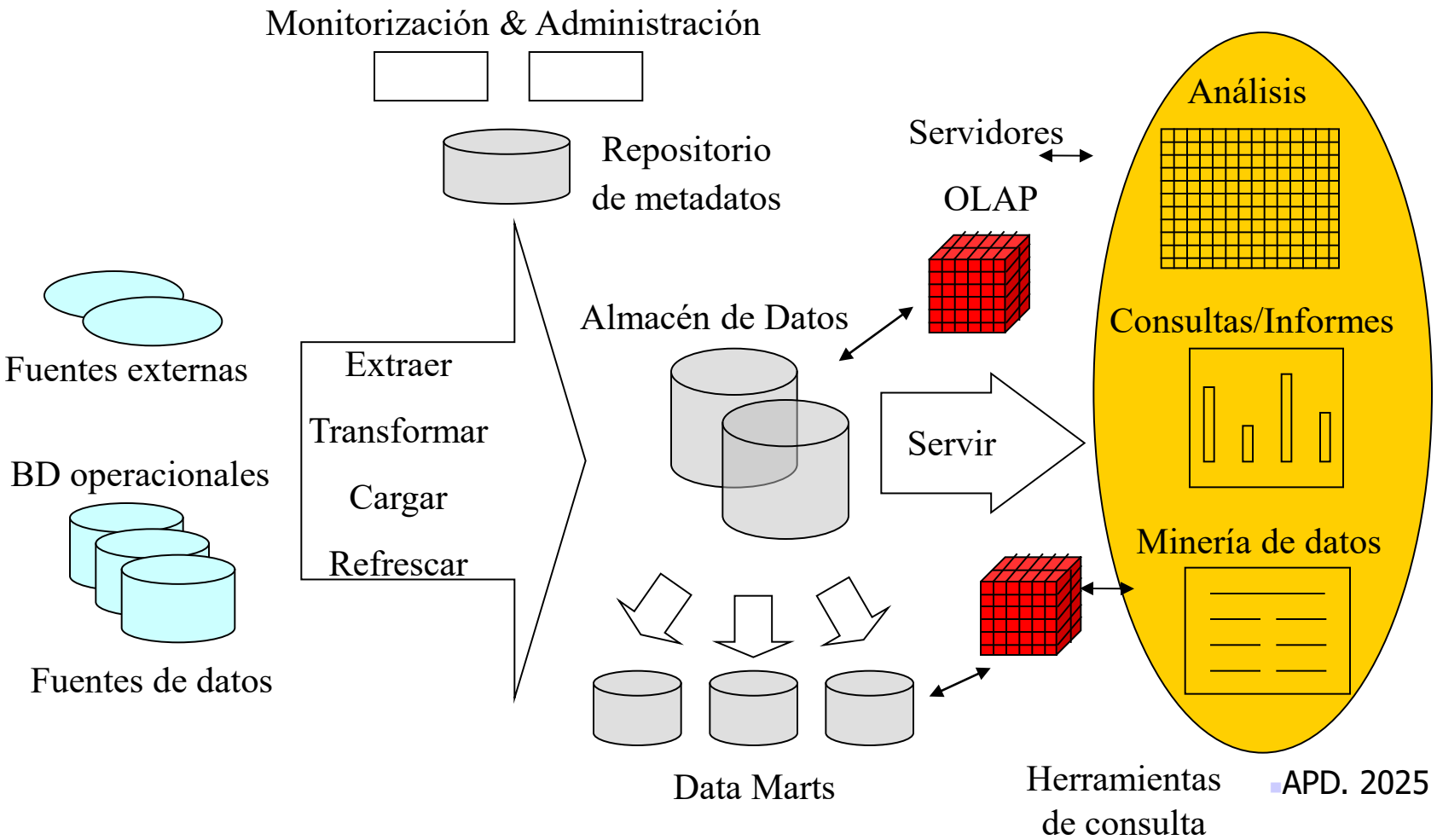
- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

Arquitecturas de DataWarehouse

- Existen distintas variantes de la arquitectura de un DW según las necesidades del negocio
- Algunas de ellas:
  - Arquitectura tradicional: Diseñada para análisis de grandes cantidades de datos estructurados
  - En tiempo real: Cuando las decisiones requieren de datos tan pronto como se generan
  - Para "Big Data": Cuando la información a tratar incluye información no estructurada (redes sociales!)

Arquitecturas de DataWarehouse

- Esquema general:



Arquitecturas de DataWarehouse

- Arquitectura de tres capas
  - Servidor del repositorio o base de datos del almacén de datos
    - Casi siempre un SGBD Relacional
  - Servidores OLAP
    - Relational OLAP (ROLAP)
      - Extiende SGBD relacionales para permitir operaciones MD
    - Multidimensional OLAP (MOLAP)

      Directamente implementa el modelo MD en vectores

Arquitecturas de DataWarehouse

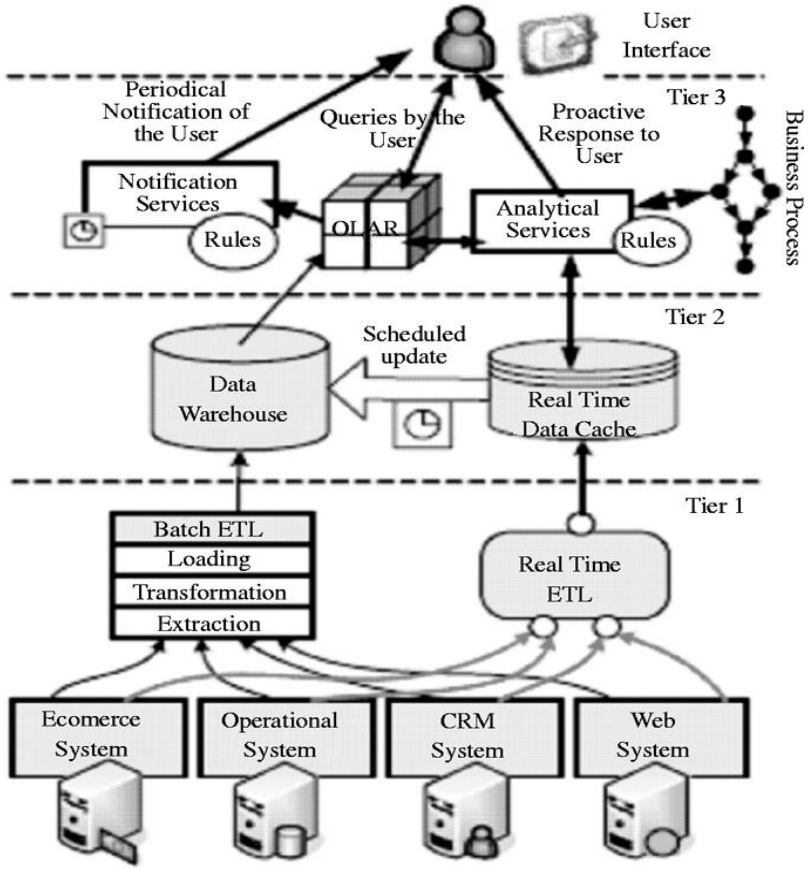
- Arquitectura de tres capas (II)
  - Clientes → Herramientas
    - Informes y consultas
    - OLAP (On-Line Analytical Processing)
    - Data Mining

Arquitecturas de DataWarehouse

- ¿Por qué Data Warehouse separado?
  - Rendimiento (Performance)
    - Consultas OLAP complejas → Ralentización del servidor
    - Métodos de implementación, accesos, etc. distintos
  - Funcionalidad
    - Datos no existentes → históricos
    - Datos consolidados (agregados, sumados, resumidos, etc.) de distintas fuentes
    - Calidad de datos
      - Diferentes fuentes → representaciones distintas, etc…

Arquitecturas de DataWarehouse

- Arquitectura tradicional: más en detalle en la siguiente sección
- En tiempo real:



Source: Nguyen Manh *et al.* (2005)

Arquitecturas de DataWarehouse

Arquitectura BigData típica

Características:

- Almacenamiento de diferentes tipos de datos
  - Semi-estructurados (Marketing/ campañas/ móbil/ web logs)
  - Estructurados
  - Ficheros de log
- Carga de datos desde diferentes bases de datos (MySQL, Oracle, PostgreSQL, MongoDB, etc)
- Minería de datos
- Analíticas
- Almacenes de datos para reporting
- Análisis por lotes (Hadoop)
- Web caching
- Search

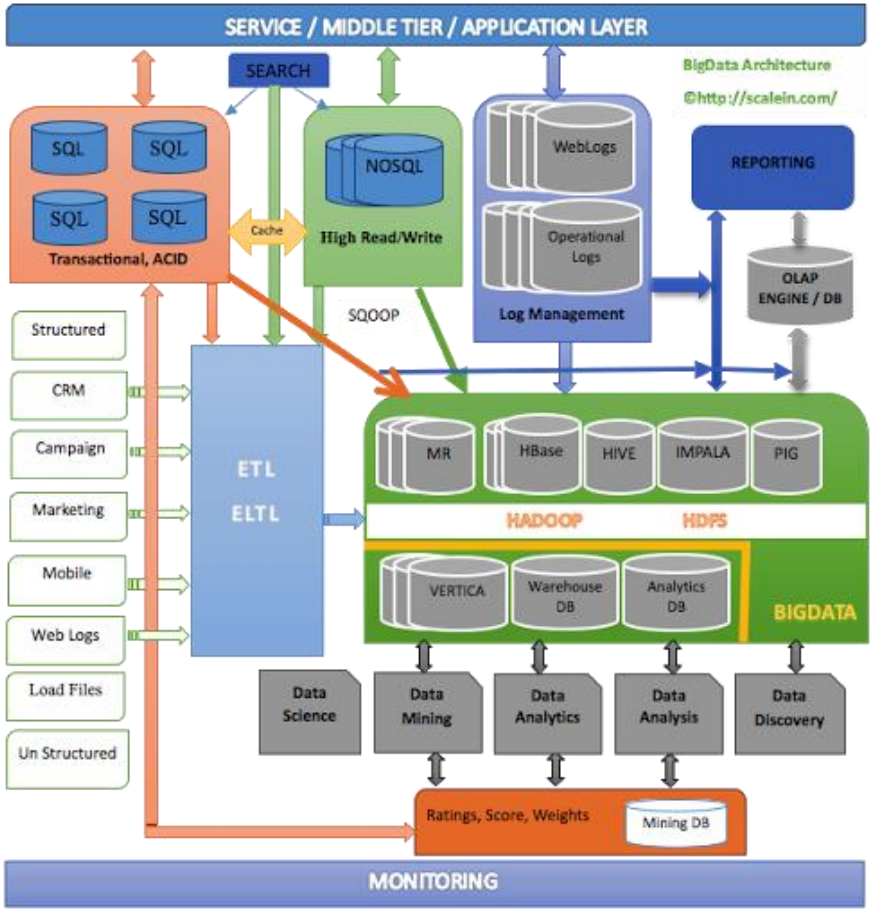


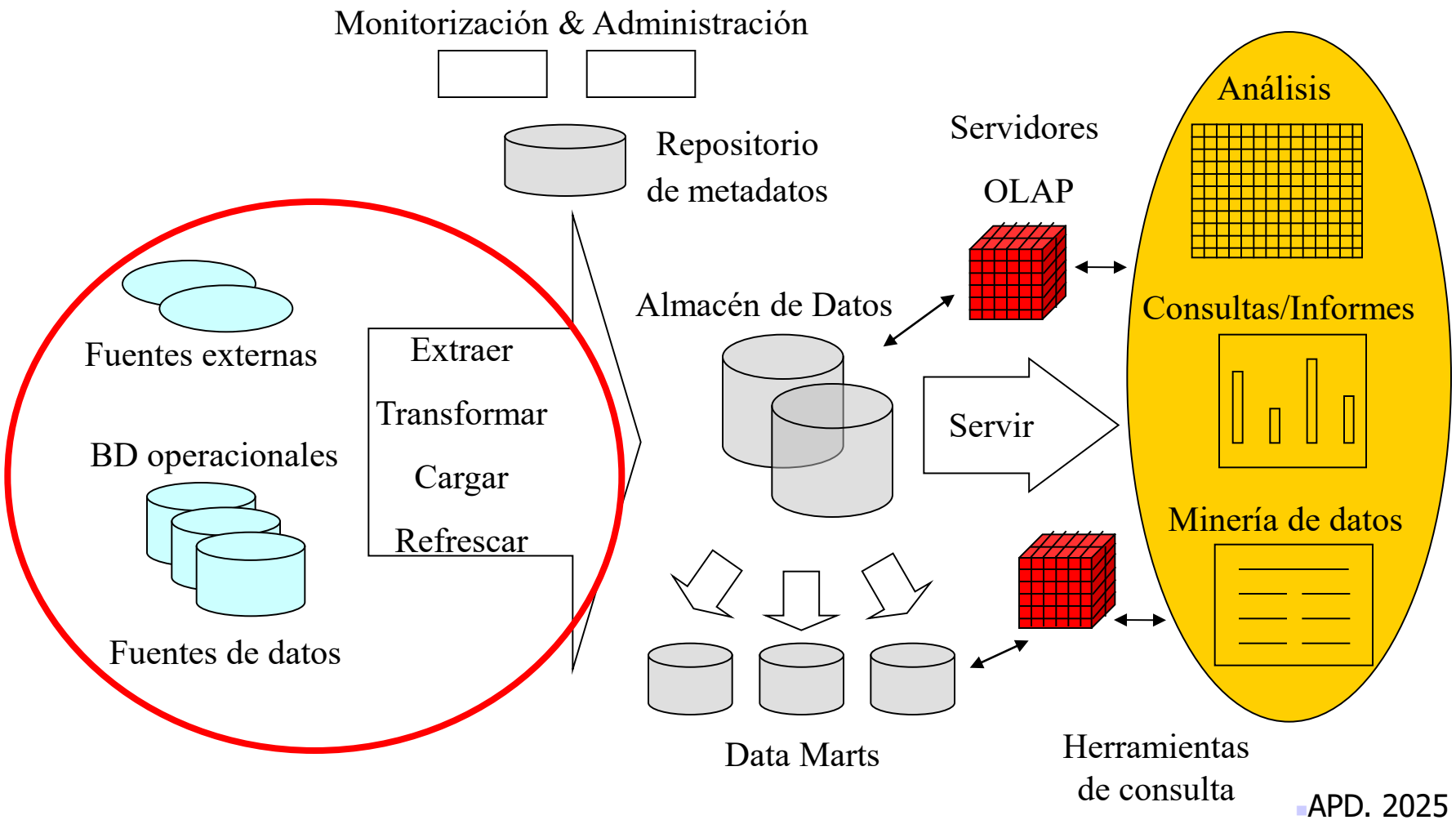
Imagen via (http://scalein.com/)

Indice

- Introducción a Big Data
- Nuevas necesidades de gestión de datos
- El almacén de datos: primera aproximación
- Diferencias entre sistemas operacionales y analíticos
- Arquitecturas de DWs
- Esquema tradicional de un DW

Esquema de una arquitectura de DW

Fuentes de datos



Esquema de una arquitectura de DW

Fuentes de datos

- Fuentes de datos
  - Fuentes de datos operacionales de la empresa
  - Bases de datos externas (públicas o privadas)
  - Ficheros planos
  - Datos en formato tradicional:
    - documentos, facturas, albaranes, etc.
  - Internet → Cantidad ingente de datos

Esquema de una arquitectura de DW

Fuentes de datos

- Procesos para alimentar de datos el almacén (ETL)
  - Extracción (Extraction)
  - Limpieza (Cleaning) y Transformación (Transformation)
  - Carga (Loading)
  - Refresco

- Extracción
  - Procesos que recogen los datos necesarios del almacén

Esquema de una arquitectura de DW

Fuentes de datos

- Limpieza (Cleaning)
  - Fundamental que los datos del almacén sean correctos
    - Decisiones estratégicas
  - Muchas fuentes de datos → alta probabilidad de error y anomalías
    - Longitud de campos inconsistentes
    - Descripción inconsistente (¿¿¿¿¿ Qué es dirección ????)
    - Valores incoherentes ("U. Alicante" vs "UA")
    - Valores nulos

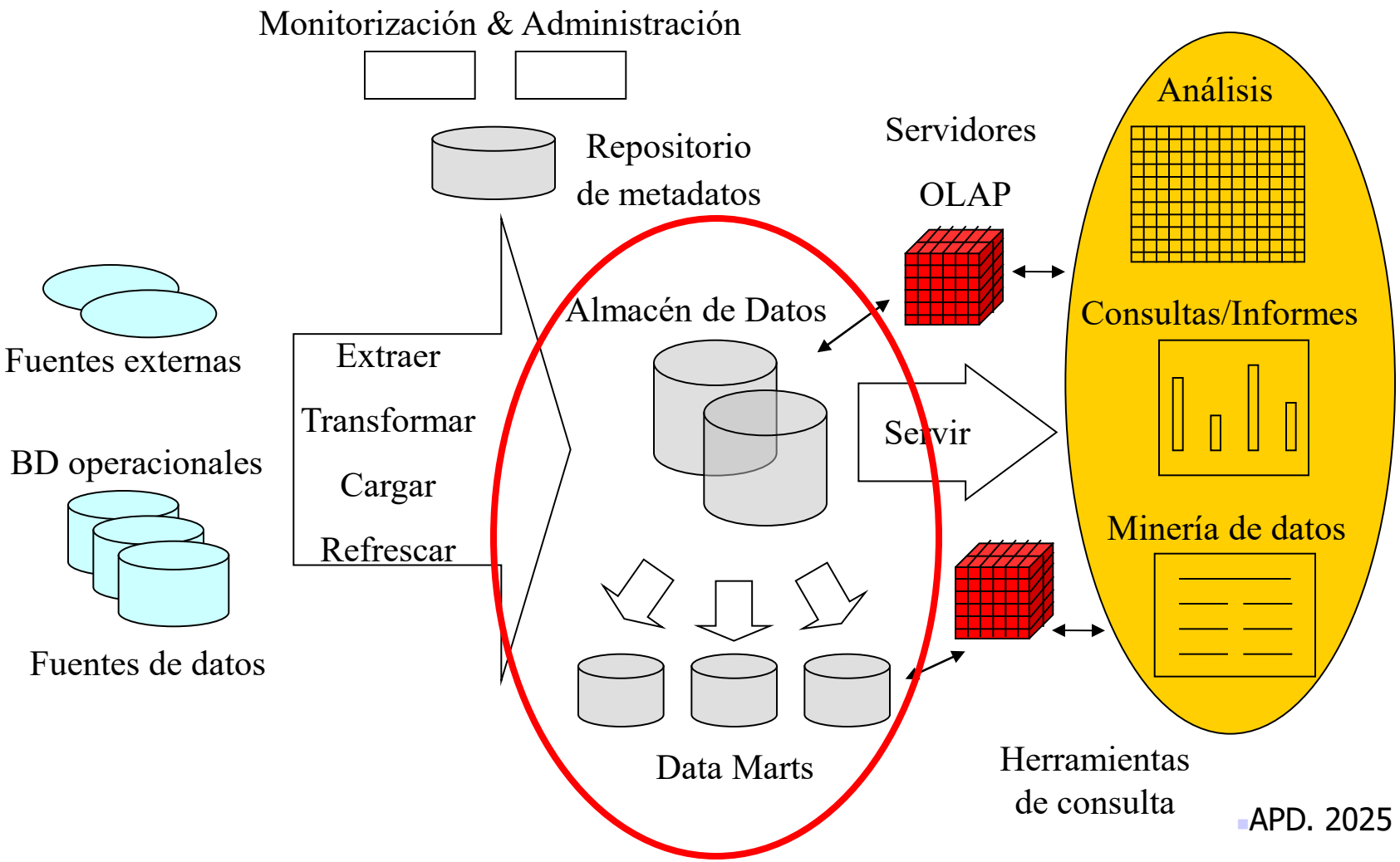
Esquema de una arquitectura de DW

Fuentes de datos

- Carga (Loading)
  - Una vez que los datos se extraen, limpian y transforman → CARGAR
  - Se puede necesitar más pre-proceso antes de carga
    - Comprobar reglas de integridad de nuevo
    - Calcular datos agregados
    - Construir Tablas derivadas y virtuales e intermedias
    - Construir índices
    - Calcular tiempo → Muchas veces por la noche
      - ¿¿ Y si multinacional ??

Esquema de una arquitectura de DW

El almacén o repositorio



Esquema de una arquitectura de DW

El almacén o repositorio

- Datos orientados por tema
  - Los sistemas OLTP están optimizados para las transacciones
    - NORMALIZAR
      - Muchas transacciones con pocos datos
        - Ej. datos de clientes en varias tablas
  - ¿Almacén de datos normalizado?: PROBLEMAS
    - ¿Directivo es capaz de tener visión de todas las tablas y relaciones?
    - Pocas transacciones que incluyen muchos datos
      - Operación MAS cara en BD: unión de tablas

Esquema de una arquitectura de DW

El almacén o repositorio

- Los datos están orientados por tema
  - En un solo lugar (digamos tabla) datos referentes a un concepto que es el objeto de estudio
    - Ej. Tabla para clientes
    - Ventas
    - Compras
    - Vehículos, etc.

Esquema de una arquitectura de DW

El almacén o repositorio

- Integrados
  - Están coherentemente agrupados a partir de datos de las fuentes de datos
  - También hay datos derivados
  - Para ello: procesos de limpieza y transformación
  - Hay errores difíciles de detectar: ¡¡¡¡¡ Cuidado !!!!!
    - Ej. Código producto válido → Exhaustivo análisis de datos

Esquema de una arquitectura de DW

El almacén o repositorio

- Integrados,.....
  - Problemas de incoherencia: resumir en 4 tipos
    - Descripción
      - J. A. Rodríguez vs. Jose A. Rodriguez
    - Codificación
      - Varón "V", Hembra "H"; en otra BD Varón "H", Hembra "M"
    - Unidades
      - Estatura: 1,70 mts; 170 cm
    - Formato
      - Número de teléfono como cadena de caracteres (965- 90 34 00) vs.como entero (965903400)

Esquema de una arquitectura de DW

El almacén o repositorio

- No volátiles
  - En sistemas OLTP se pueden modificar datos (ej. tuplas)
    - Unidades de pedido 200; si cliente modifica, se cambia.
  - En DW nunca se modifican (salvo excepciones), se añaden nuevos datos para el análisis
    - Un pedido con una fecha de 200 uds.
    - Un pedido con otra fecha de 150 uds.

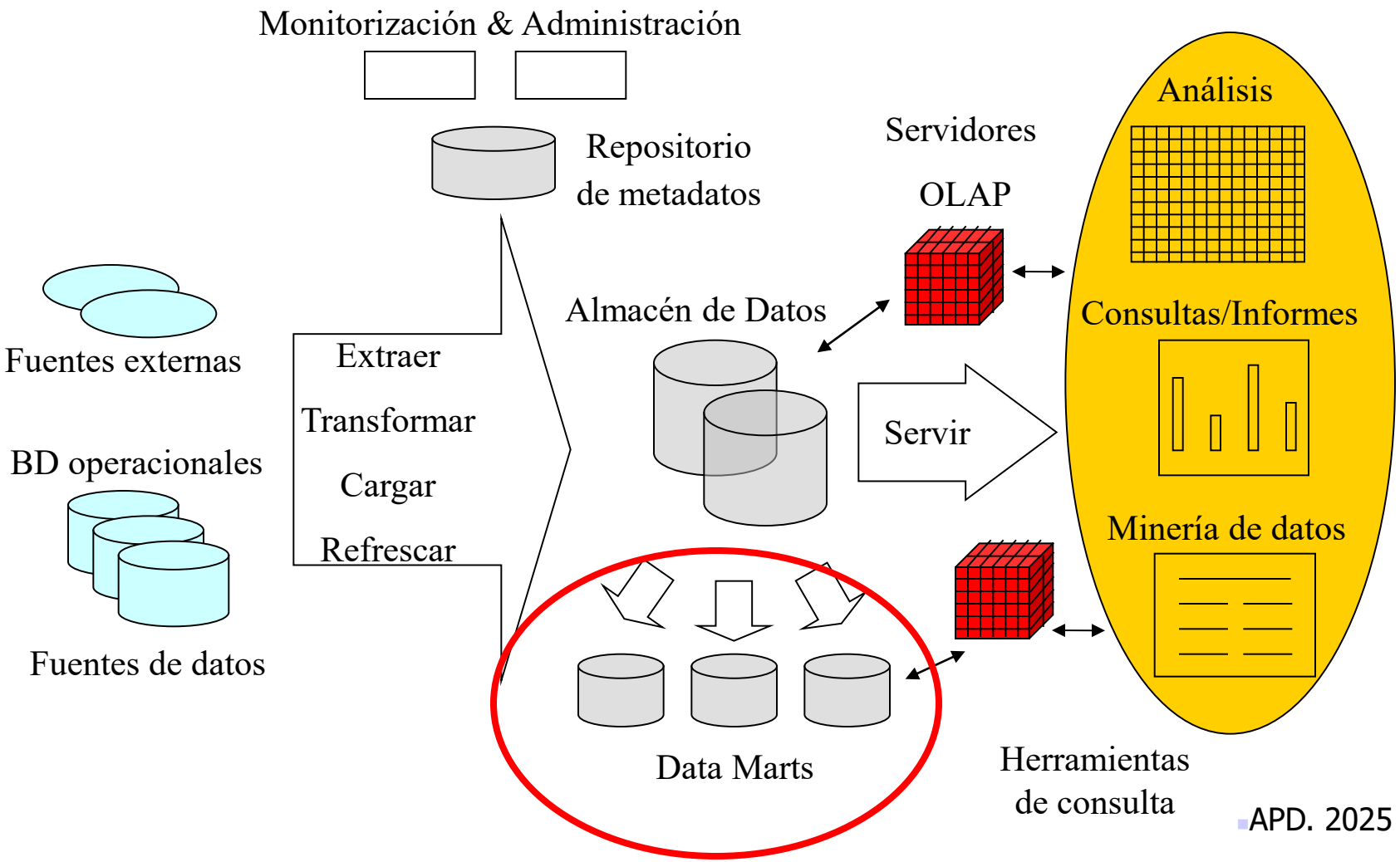
Esquema de una arquitectura de DW

El almacén o repositorio

- Variables en el tiempo
  - No volatibilidad → Dimensión básica: El TIEMPO
    - Datos analizados en función del tiempo
    - Ej. Anterior del pedido
      - ¿ Por qué un cliente ha variado la cantidad de su pedido en una semana ?
      - ¿ Por qué han cambiado los gustos de un cliente en varios años ?
      - Etc.

Esquema de una arquitectura de DW

El almacén o repositorio



Esquema de una arquitectura de DW

El almacén o repositorio

- Data Marts
  - Es como una vista del almacén de datos
  - Se definen para satisfacer las necesidades de un departamento o sección dentro de una empresa
  - Normalmente, en la práctica, suelen contener más cantidad de información agrupada que en detalle

Esquema de una arquitectura de DW

El almacén o repositorio

- Data Marts.....
  - Para su construcción se pueden seguir dos aproximaciones:
    - (I) Definir primero el almacén de datos y, a partir de él, definir los data marts, ó
    - (II) Definir primero los data marts departamentales y, posteriormente integrarlos en un almacén de datos global para la organización
  - Nota: Si la envergadura de la empresa es considerable y, la experiencia en construir DW poca, es aconsejable seguir la aproximación 2

Esquema de una arquitectura de DW

El almacén o repositorio

- Resumen: Data warehouse vs. Data Marts
  - Enterprise DW
    - Información sobre "temas" de toda organización
    - Requiere complejo modelado de negocio
    - Puede llevar AÑOS para construir e implementar
  - Data Mart
    - Departamental → sub-temas
      - Ej. Marketing data mart, Clientes, productos, ventas !!!!!
    - Más rápido agregar
    - OJO !!! Integración con DW puede ser compleja

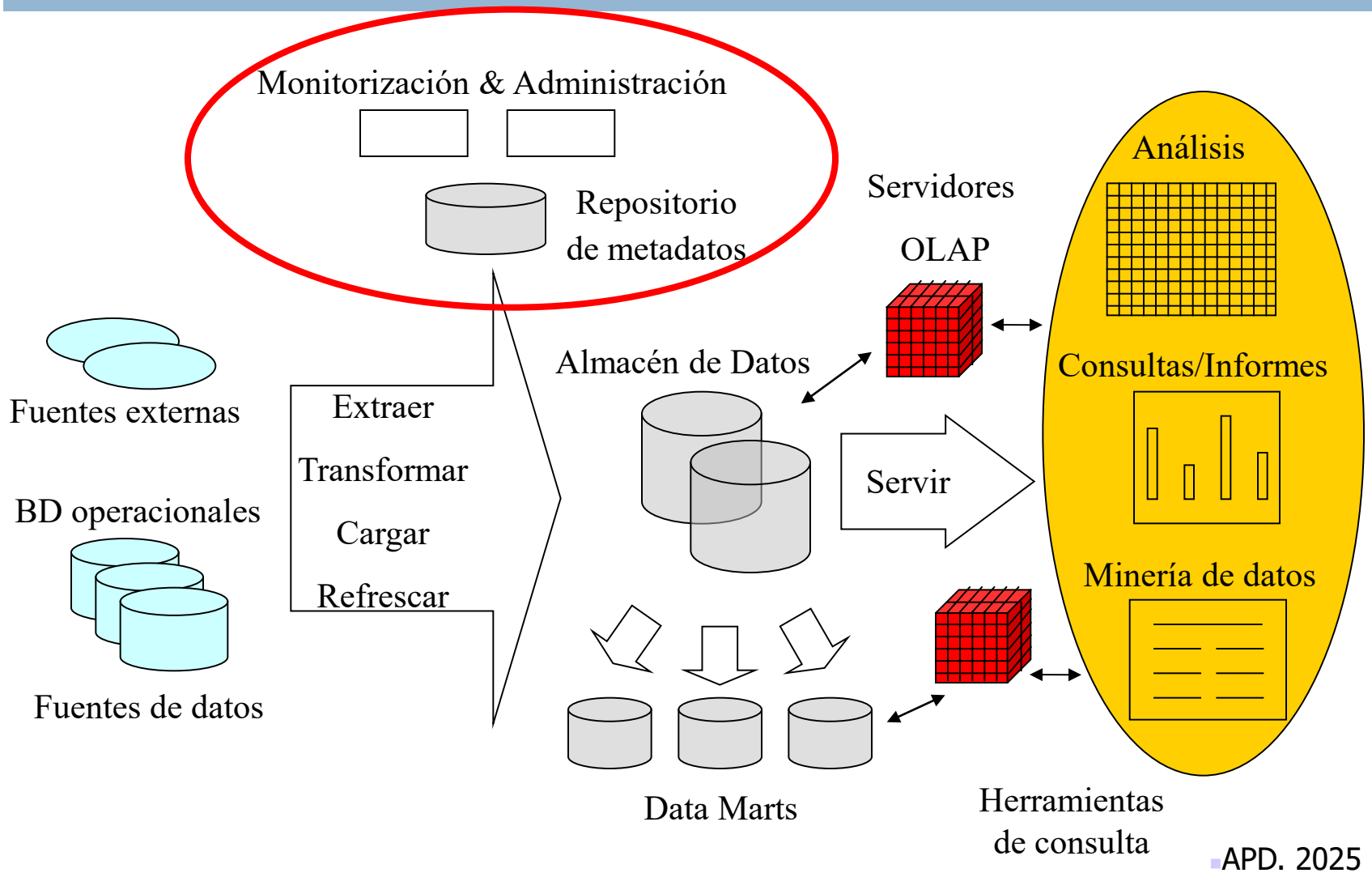
Esquema de una arquitectura de DW

El almacén o repositorio

- Virtual Data Warehouse
  - Vistas sobre bases de datos operacionales
    - Materializan algunas vistas agregadas → consultas eficientes
    - Mayor facilidad en la construcción
    - Requieren exceso de capacidad del servidor operacional
    - Muchas empresas dicen tener DW cuando el administrador ha creado un VIRTUAL DW

Esquema de una arquitectura de DW

Los metadatos



Esquema de una arquitectura de DW

Los metadatos

- Los metadatos
  - Son datos sobre datos
    - Qué dato se guarda (ej. clientes)
    - Dónde se guarda (tabla clientes)
    - Campos de la tabla
    - Con qué datos de las fuentes se corresponden
    - Niveles de agregación
    - Procesos de carga → ¿ Cuándo se actualizan ?
      - ¿ Cuándo fue la última actualización?
    - Patrón de dato válido (Ej. *Apellido1 Apellido2, Nombre*)

Esquema de una arquitectura de DW

Los metadatos

- Los metadatos …
  - Son datos sobre datos …
    - Reglas de transformación
    - ¿ Cuándo se incorporan al almacén de datos ?
    - Y muchos más…

Esquema de una arquitectura de DW

Los metadatos

- Tipos
  - Administrative metadata (Toda la información necesaria para el DW)
    - Fuentes de datos y contenidos
    - Esquema del data warehouse, vistas y datos agregados
    - Dimensiones de análisis con sus jerarquías
    - Consultas e informes predefinidos
    - Localización y contenido del los Data Marts
    - Diseño físico → particionamiento de datos

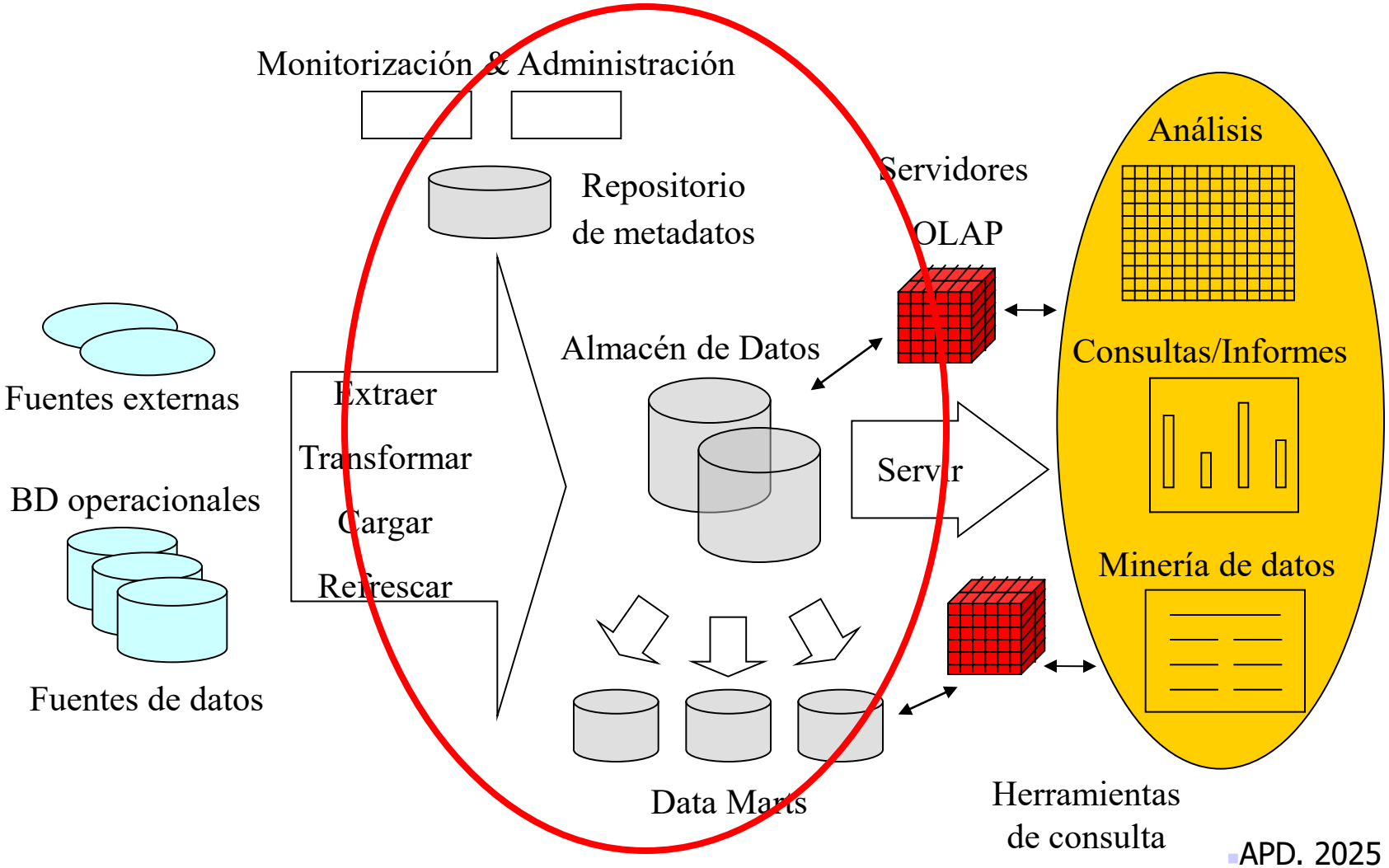
Esquema de una arquitectura de DW

Los metadatos

- Tipos…
  - Business metadata
    - Información y términos de negocio
    - Políticas de posesión de datos
    - Políticas de permiso de datos por usuarios (seguridad)
  - Operational metadata
    - Obtiene información recogida durante el proceso del "almacén de datos"
      - Datos migrados y secuencia de transformaciones aplicadas
      - Auditoría
      - Informes de error

Esquema de una arquitectura de DW

Servidor del almacén de datos



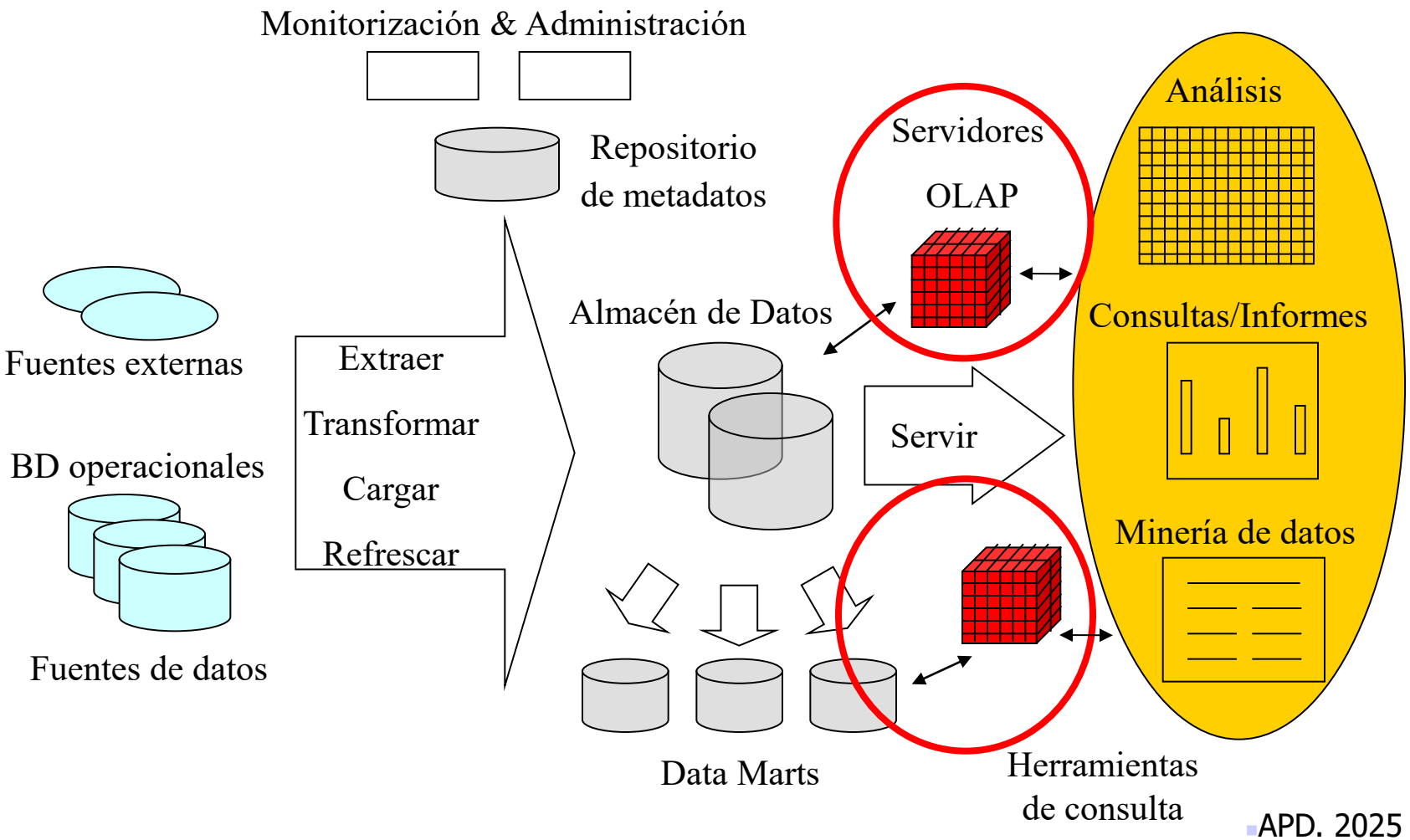
Esquema de una arquitectura de DW

Servidor del almacén de datos

- El servidor es un SGBD que se encarga de
  - Gestionar el repositorio propio del almacén de datos
  - Coordinar los procesos ETL que alimentan el DW
  - Procesan las consultan lanzadas sobre el almacén y devuelven los datos
  - Generalmente son servidores relacionales

Esquema de una arquitectura de DW

Servidor del almacén de datos



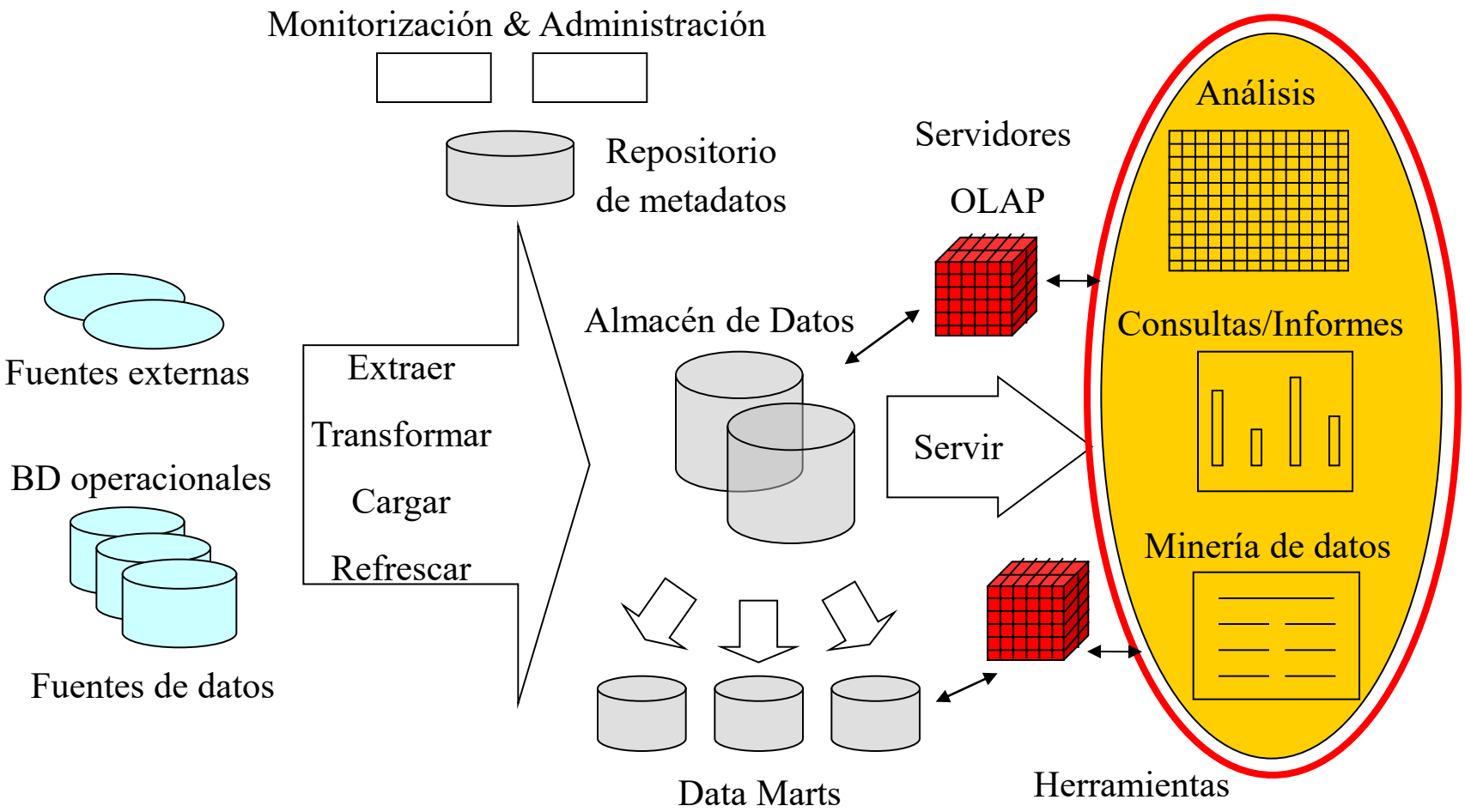
Esquema de una arquitectura de DW

Servidor de consultas

- El servidor de consultas
  - En la mayoría de las arquitecturas se utiliza un servidor distinto al del almacén de datos
    - Rendimiento y mantenimiento
  - La mayoría de herramientas funcionan con esta arquitectura
    - Ejplo. MicroStrategy
  - Dos tecnologías ampliamente utilizadas
    - ROLAP
    - MOLAP

Esquema de una arquitectura de DW

Herramientas de consultas



Esquema de una arquitectura de DW

Herramientas de consultas

- Generación de informes
  - Consultas ad-hoc e informes
    - Permiten acceso a los datos base
    - Informe se construye con “point-and-click”
  - Entornos consulta/informes de directivos
    - Muestra datos base en términos de negocio
    - Utiliza consultas predefinidas/almacenadas
    - Soporte limitado para consultas ad-hoc
    - Informes se pueden presentar como documentos

Esquema de una arquitectura de DW

Herramientas de consultas

- Data Mining (Minería de datos)
  - Descubre tendencias y patrones (minería interpretativa)
  - Crean modelos y hacen predicciones (minería predictiva)
    - Utilizan los datos existentes para detectar tendencias y crear modelos
    - Aplican los patrones y modelos a datos nuevos
    - Importante → Adaptación de modelos según nuevos datos

El almacén de datos

Bibliografía

- Juan Trujillo, Jose Norberto Mazón, Jesús Pardillo. Diseño y explotación de almacenes de datos: Conceptos Básicos de Modelado Multidimensional. Ed. Club Universitario. 2011.
  - Tema 1
- Inmon (2005). Building the Data Warehouse (4ª ed.)
  - Tema 1 (introducción)
- Kimball & Ross (2013). The Data Warehouse Toolkit (3ª ed.)
  - Tema 1 (introducción)
- Thomsen (2000). OLAP solutions: Building Multidimensional Information Systems
  - Tema 1 (introducción)
- Giovinnazo (2000). Object-Oriented Data Warehouse Design: Building a star schema
  - Tema 1 (introducción)
- Barrera, J. M., Reina-Reina, A., Lavalle, A., Maté, A., & Trujillo, J. (2024). An extension of iStar for Machine Learning requirements by following the PRISE methodology. *Computer Standards & Interfaces*, 88, 103806.
- Lavalle, A., Reina-Reina, A., Maté, A., Trujillo, J. (2024). Social Modeling Using the i* Framework.
  - Capitulo: Data Analytics from a Social Perspective.

Introducción al concepto de los datos y su análisis

Tema 1

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

DISEÑO CONCEPTUAL DE ALMACENES DE DATOS

Tema 2

Profesores:

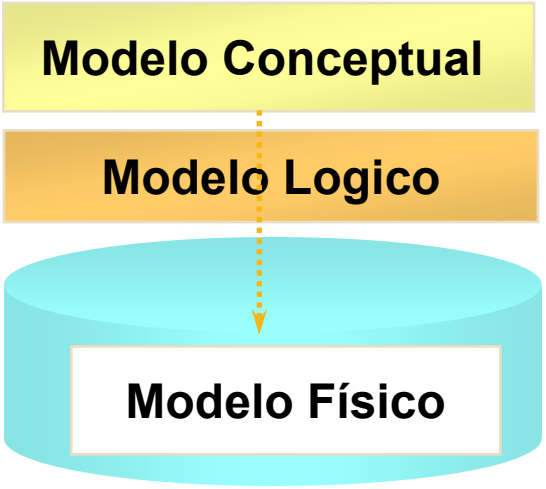
Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

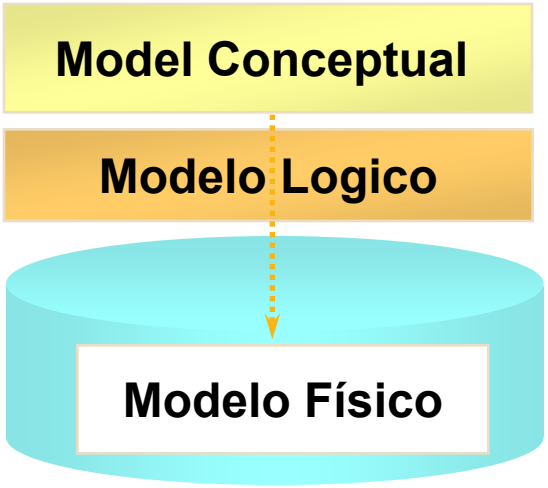
Departamento de Lenguajes y Sistemas Informáticos

Modelos utilizados en BD

- Modelo Conceptual
- Modelo Logico
- Modelo Físico



Sistemas Operacionales

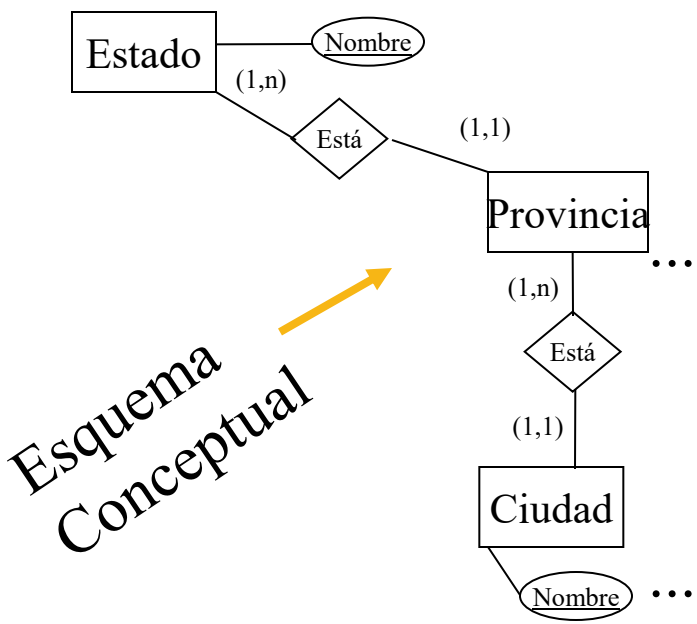


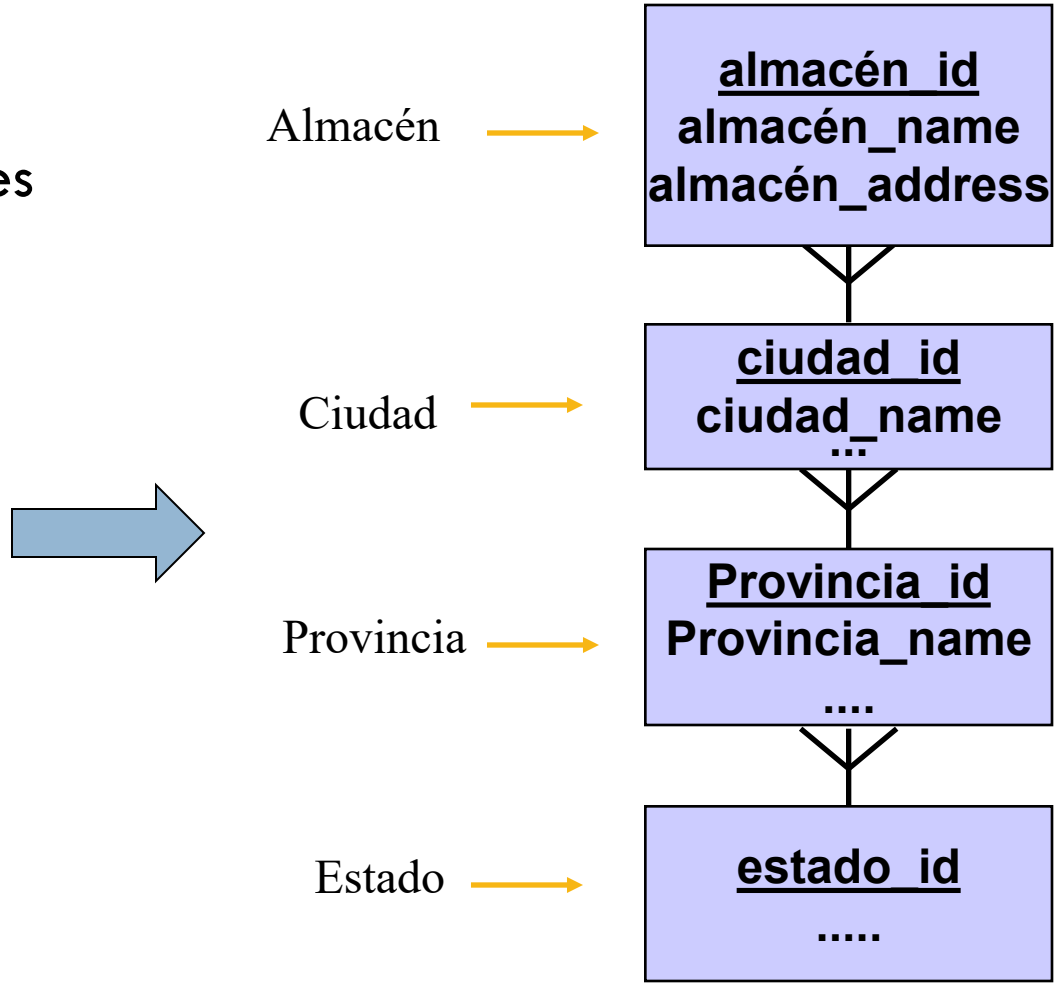
Almacenes de Datos

Modelos utilizados en BD

Sistemas operacionales

- Normalización
  - Sist. Transaccionales





Diseño conceptual de DW

- Modelado multidimensional (MD)
  - Parte estructural
  - Parte dinámica
- Parte estructural
  - Hechos y dimensiones
- Parte dinámica
  - Operaciones de consulta al modelo MD

Modelado multidimensional

- Tal y como el usuario percibe el mundo real objeto de estudio
  - Perspectiva estructural
    - Modelado o Modelo Multidimensional (MD)
      - Hechos y Dimensiones
  - Perspectiva dinámica
    - Definición de requerimientos iniciales sobre el modelo MD
    - Operaciones de consulta avanzada

Modelado multidimensional

Parte estructural

- Perspectiva estructural → modelo MD
  - Hechos
    - Objeto de análisis
      - Ej. Ventas de productos, compras, alquileres, transportes
  - Dimensiones
    - Diferentes perspectivas para analizar los hechos
      - Ej. Productos, almacenes, tiempo, vehículos, etc.

Modelado multidimensional

Parte estructural

- Hechos representan normalmente relaciones *muchos a muchos* con todas las dimensiones y, *muchos a uno* con cada dimensión en particular
  - Ej. Ventas de productos (H) por producto (D), almacenes (D) y tiempo (D)
    - Un producto (D) → varias ventas (H)
    - Una venta (H) → un solo producto (D) y almacén (D)

Modelado multidimensional

Parte estructural

- Sin embargo, a veces hechos son *muchos a muchos* con dimensiones en particular
  - Ej. Tickets emitidos (H) por ....
    - Un ticket (H) puede contener muchos productos
- Hechos y Dimensiones se caracterizan por atributos
  - Hechos → atributos de hecho o medidas
  - Dimensiones → atributos de dimensión

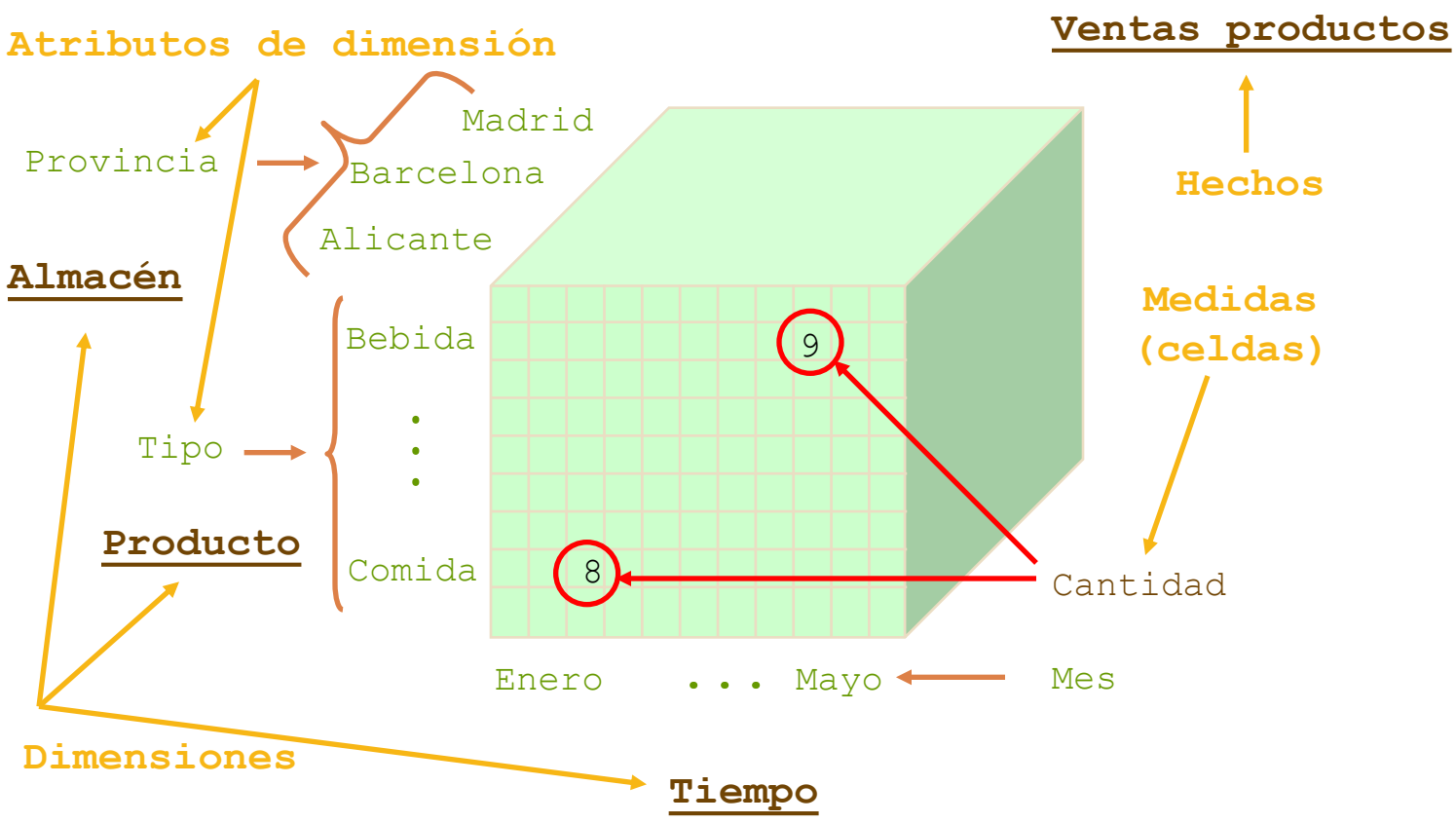
Modelado multidimensional

Parte estructural

- ¿Cómo se representa el modelo MD intuitivamente?
  - Cubos
  - Hipercubos (Cubos sobre cubos)
  - Tablas multidimensionales tipo hoja de cálculo, etc.

Modelado multidimensional

Parte estructural. Cubo



Modelado multidimensional

Parte estructural. Tablas MD.

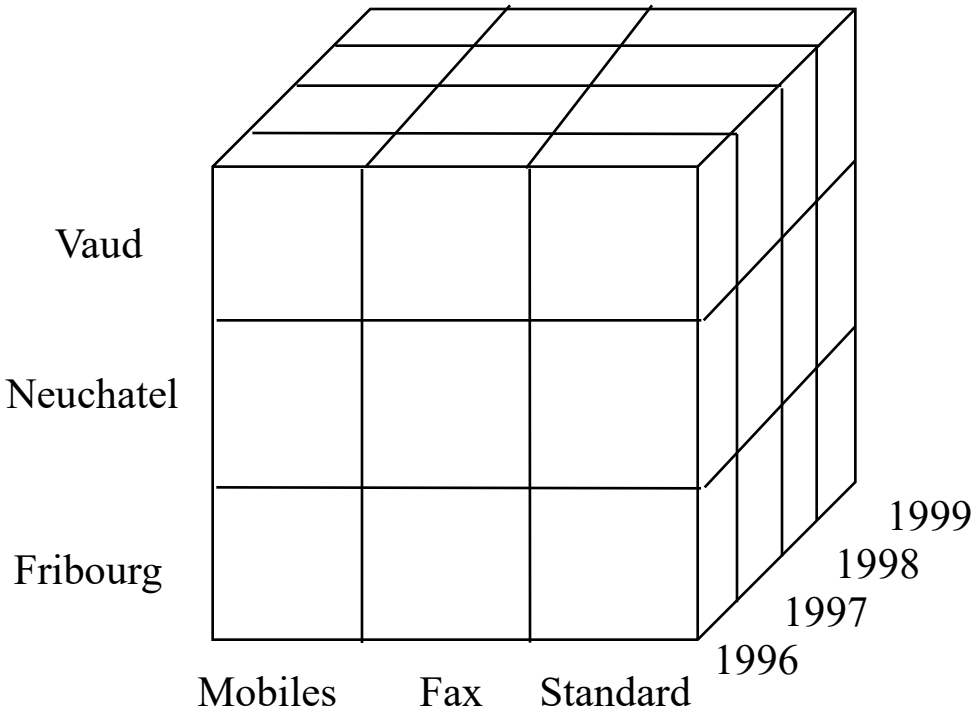
- Tablas multidimensionales

Ventas			Producto.Grupo = "Supermercado"			
			Comida		Bebida	
			Cong	Fresco	Refresco	Alcohol
Almacén. comunidad = "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Burjasot	900	1000	1100	1200
		Cullera	1300	1400	1500	1600

Modelado multidimensional

Parte estructural. Cubo vs. Tabla relacional.

- Para recuperar los datos necesarios:
  - En una SGBDR: 4*3*3 = 36 tuplas (filas)
  - En una BD Multidimensional: 4+3+3 = 10 valores en los ejes



Region	Producto	Año	Llam	*tupla*
Vaud	Fax	1997	12	*1*
Vaud	Mobiles	1998	23	*2*
Vaud	Standard	1999	22	*3*
Fribourg	Fax	1997	34	*4*
Fribourg	Mobiles	1998	45	*5*
Fribourg	Standard	1999	48	*6*
Neuchatel	Fax	1986	55	*7*
Neuchatel	Mobiles	1987	66	*8*
...	...	...	...	...
Neuchatel	Standard	1998	55	*35*
Neuchatel	Mobiles	1999	66	*36*

Modelado multidimensional

Parte estructural. Dimensiones

- Puede haber alto grado de categorización
  - Atributos en función de instancias
    - Ej. Volumen y porcentaje de alcohol sólo para bebidas
    - Ej. Tiempo y modo preparación sólo para comidas
- Atributos dimensión → jerarquías clasificación
  - Los niveles de jerarquía serán usados para la agregación de las medidas
    - Ej. Ciudad, comunidad, tipos de productos, etc.

Modelado multidimensional

Parte estructural. Dimensiones. Jerarquías de clasificación

- Las instancias de niveles ≡ miembros

- Clasificación
  - Cardinalidad
    - Por defecto → estrictas (*1-m*)
      - Una instancia sólo se relaciona con una instancia del nivel superior de jerarquía
        - Ej. Un almacén está ubicado en una sola ciudad
    - Sin embargo algunas pueden ser **no estrictas (m-m)**
      - Ej. Un almacén pertenece a más de una zona de ventas

Modelado multidimensional

Parte estructural. Dimensiones. Jerarquías de clasificación

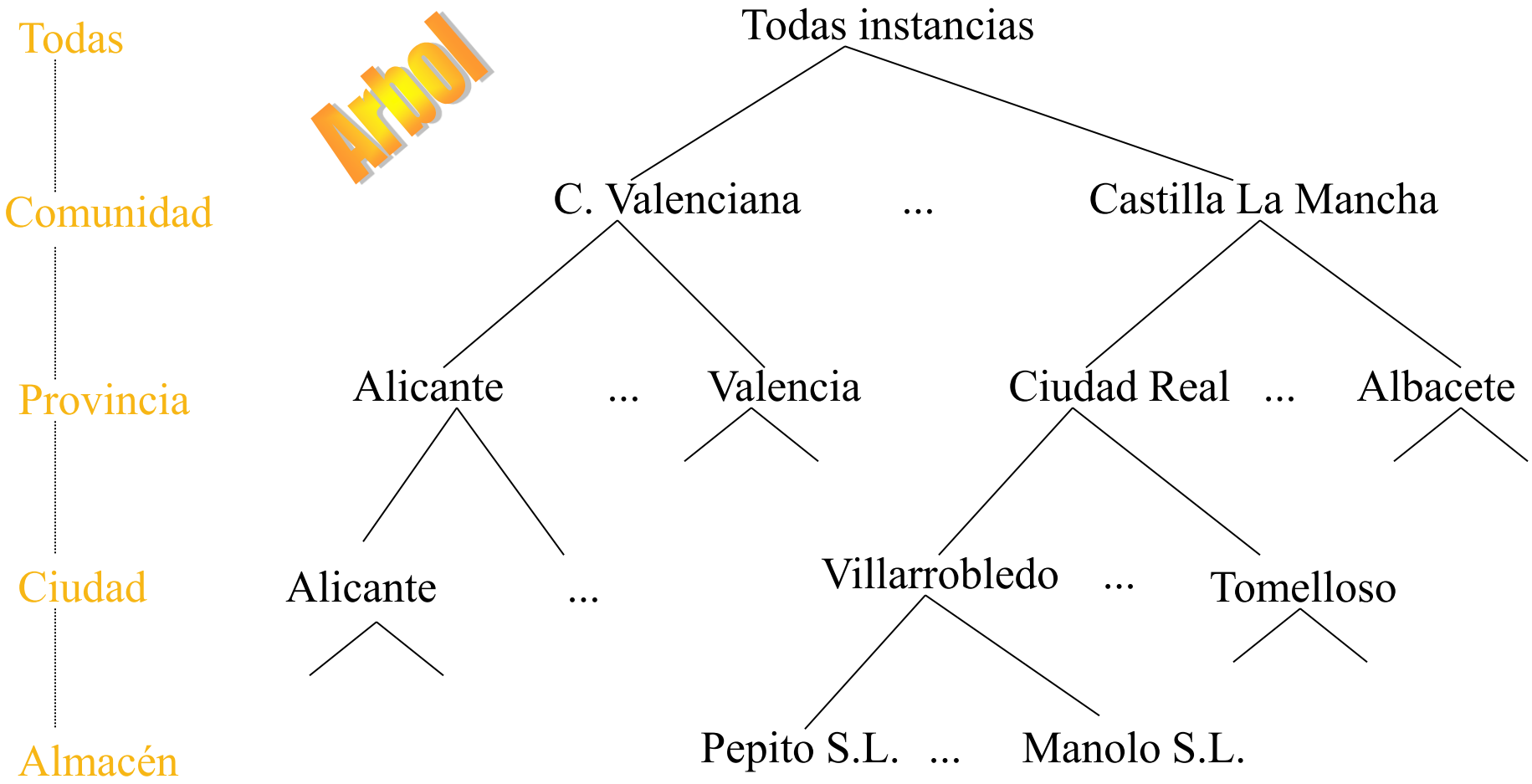
- Caminos
  - Simples → Representación mediante árbol

Simétricas →	Todos los niveles de la jerarquía son obligatorios y deben estar instanciados.
Asimétricas →	Los niveles inferiores de la jerarquía no son obligatorios y pueden no estar instanciados.
Generalizadas →	Camino de navegación establecido en función de una categorización de instancias.

  - Múltiples → Representación mediante grafo

Modelado multidimensional

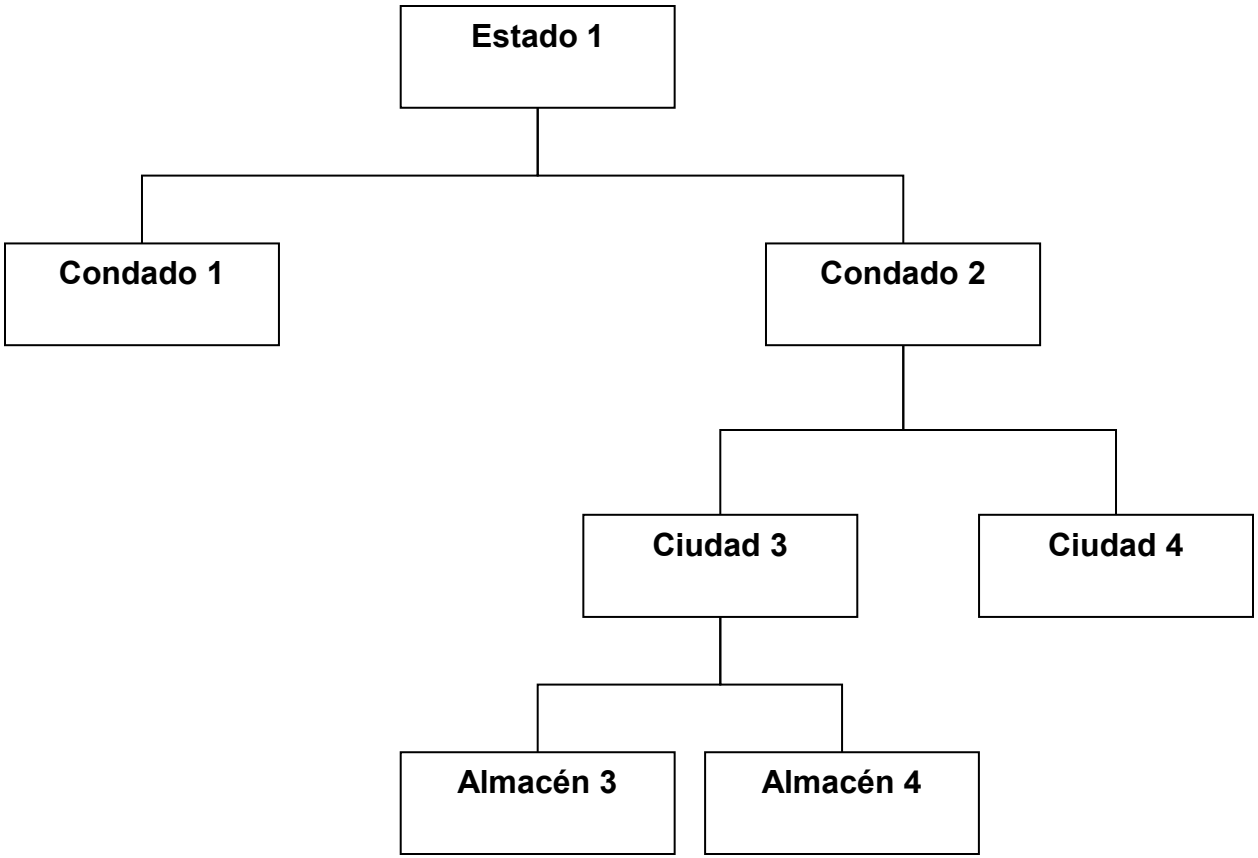
Parte estructural. Dimensiones. Jerarquías de clasificación simples.



Modelado multidimensional

Parte estructural. Dimensiones. Jerarquías de clasificación simples.

- Ejemplo de Asimétrica



Modelado multidimensional

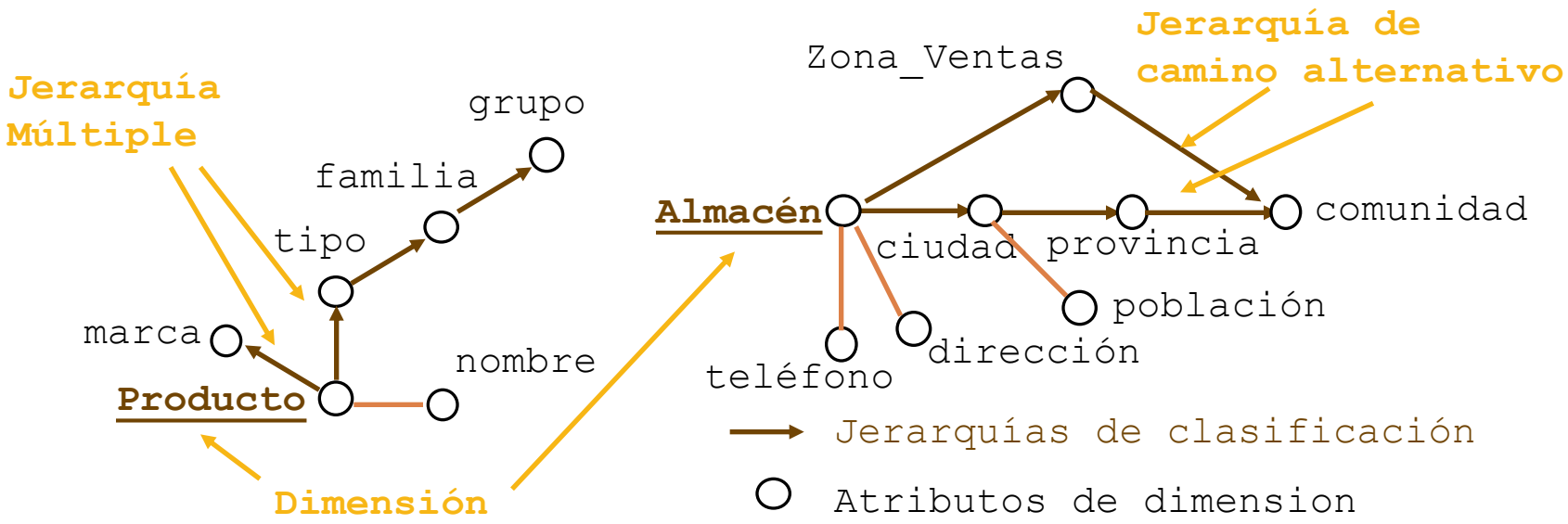
Parte estructural. Dimensiones. Jerarquías de clasificación

- Jerarquías múltiples y de camino alternativo
  - Rep. → D.A.G.
    - Ej. Ciudad se puede clasificar en comunidad y,
    - Ciudad también se puede clasificar en zona ventas
- Jerarquías paralelas → Más de una jerarquía definida para la misma dimensión
  - Independiente → Las distintas jerarquías no comparten niveles
  - Dependientes → Las distintas jerarquías comparten algún nivel

Modelado multidimensional

Parte estructural. Dimensiones. Jerarquías de clasificación

- Multiples → Normalmente se representan mediante Grafos Acíclicos dirigidos (G.A.D.)



Modelado multidimensional

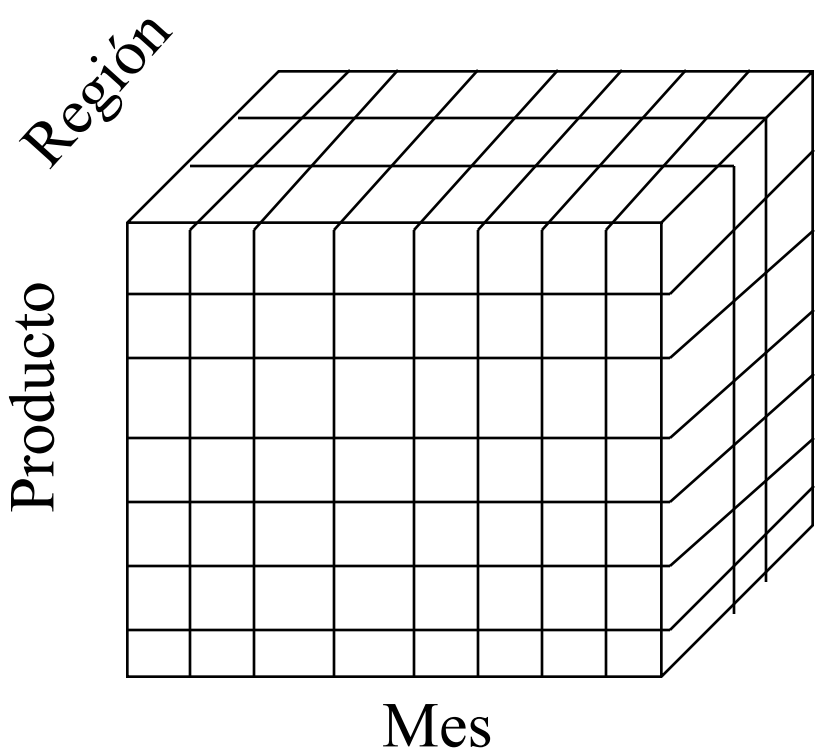
Parte estructural. Dimensiones. Jerarquías de clasificación

- Jerarquías completas
  - Además de estrictas, un miembro o instancia de un nivel superior está compuesto únicamente por los del nivel inferior → relación fija entre instancias

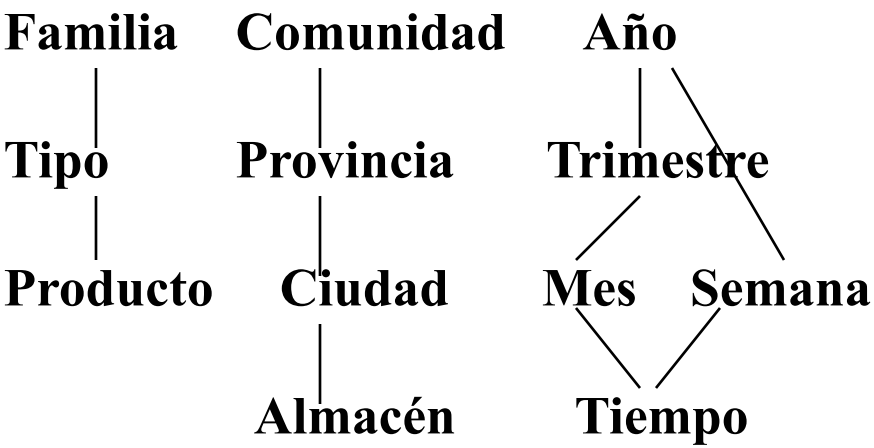
Modelado multidimensional

Parte estructural. Dimensiones. Jerarquías de clasificación y cubos

- Si utilizamos un cierto nivel de agregación
  - Ventas es una función del producto, mes y región

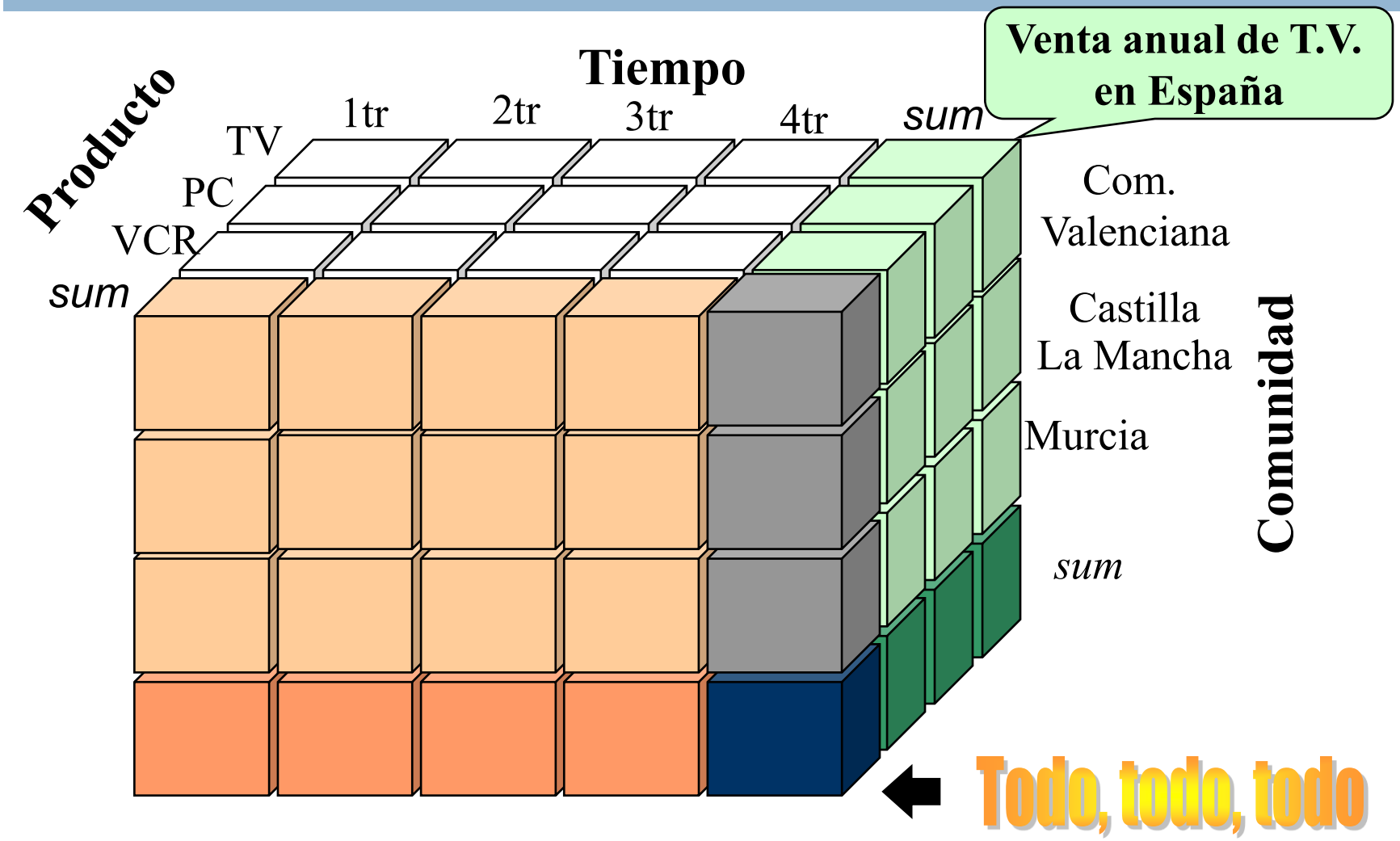


Dimensiones: Producto, Almacén, Tiempo
Caminos de jerarquía por los que agregar



Modelado multidimensional

Parte estructural. Un ejemplo de cubo de datos



Modelado multidimensional

Parte estructural. Hechos

- Atributos de hecho o medidas
  - Atómicos
    - Ej. Cantidad vendida, precio, etc.
  - Derivados
    - Utilizan una fórmula para calcularlos
    - Ej. Precio_total = precio * cantidad_vendida

Modelado multidimensional

Parte estructural. Hechos

- Aditividad
  - Conjunto de operadores de agregación (SUM, AVG, etc.) que se pueden aplicar para agregar los valores de medidas a lo largo de las jerarquías de clasificación *(Kimball, 1996)*
  - Es aditiva → SUM sobre todas las dimensiones
  - Semi-aditiva → SUM sólo sobre algunas dimensiones
  - No aditiva → SUM sobre ninguna dimensión

Modelado multidimensional

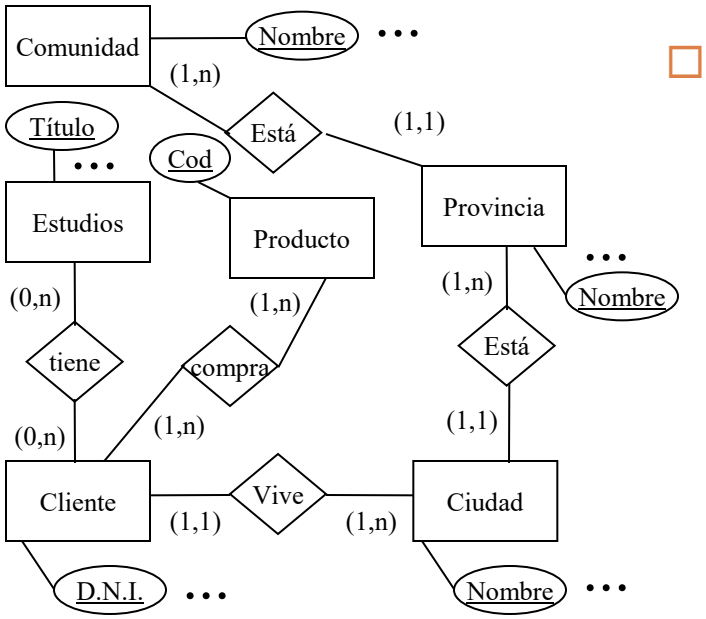
Parte estructural. Hechos

- Si no aditiva → otros operadores pueden aplicarse (ej. AVG, MIN, etc.)
  - Ej. Atributos que miden niveles (ej. Inventarios) no son aditivos sobre la dimensión tiempo
    - Es aditivo sobre la dimensión producto
  - Las medidas de temperatura no son aditivas
  - Algunas son semánticamente incorrectas
    - Ej. Atributo número de clientes que cuenta el número de tickets emitidos no es aditiva sobre la dimensión producto

Modelado multidimensional

- En aplicaciones OLTP…
  - Modelado conceptual → Entidad-Relación (EER)
    - ¿ Podría reflejar la multidimensionalidad de los datos ?
    - Hechos,…
    - Dimensiones,…
  - ¿ Podría ser interrogado por un analista de la información ?

Modelado multidimensional



- ¿Hechos?
- ¿Dimensiones?
  - ¿Jerarquías?

Transformar

- ¿Podría un analista interrogarlo?

Modelado multidimensional

Parte dinámica.

- BD Multidimensionales parecen más naturales y
- ⇒ Queries son también más naturales
- Ejemplo: llamadas de tfno. por producto y región
  - En un cubo: *print total.(calls keep product, region)*

Producto/Comun.	C. Valenciana	Cast. Mancha
Fax	44	28
E-mail	27	51
Mobiles	46	11

  - En tablas relacionales:

Select producto, comunidad, sum(llamadas)
from Llamadas, Comunidad
where llamadas.comunidad= Comunidad.comunidad
Group by producto, comunidad
Order by producto, comunidad;

Product	Region	Sum(calls)
Fax	Vaud	44
Fax	Valais	28
Mobile	Vaud	27
Mobile	Valais	51
Standard	Vaud	46
Standard	Valais	11

Modelado multidimensional

Parte dinámica.

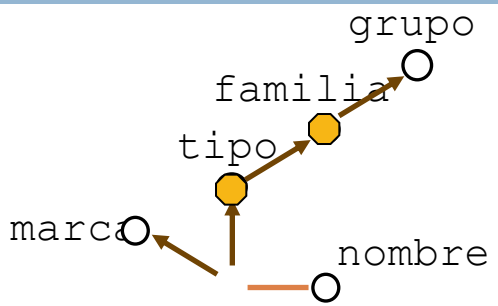
- Definición de requerimientos iniciales de usuario
  - Están basados en jerarquías definidas en Dimensiones

Cantidad vendida de productos comestibles agrupados por su familia y tipo, vendidos en la comunidad valenciana y, agrupados por la provincia y ciudad donde se vendieron

Modelado multidimensional

Parte dinámica.

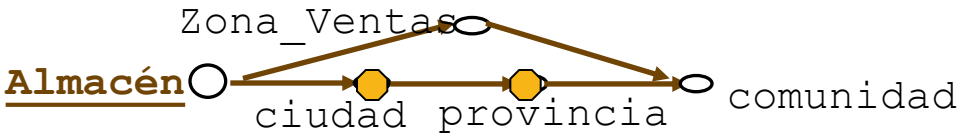
Operaciones de consulta (OLAP)



Roll-up

- Agregar valores de medidas a lo largo de jerarquías de clasificación

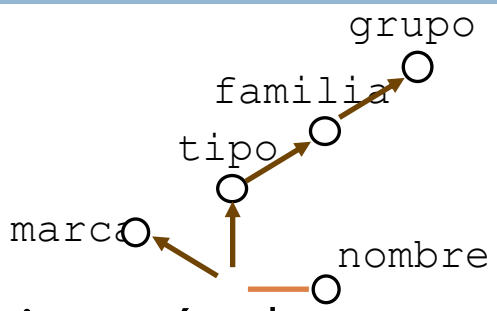
Ventas'		Producto.Grupo = "alimentación"	
		Comida	Bebida
Almacén. comunidad = "Comunidad Valenciana"	Alicante	1400	2200
	Valencia	4600	5400



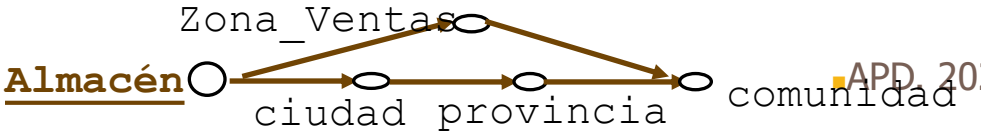
Modelado multidimensional

Parte dinámica.

- Operaciones de consulta (OLAP)
  - Drill-down
    - Desagregar valores de medidas a lo largo de jerarquías de clasificación



Ventas			Producto.Grupo = "alimentación"			
			Comida		Bebida	
			Cong.	Fresca	Refresco	Alcohol
Almacén. Comunidad= "Comunidad Valenciana"	Alicante	Albatera	100	200	300	400
		Elche	500	600	700	800
	Valencia	Burjasot	900	1000	1100	1200
		Cullera	1300	1400	1500	1600



Modelado multidimensional

Parte dinámica.

- Operaciones de consulta (OLAP)
  - Drill-accross
    - Consultar medidas de varios hechos en el mismo cubo
      - Ej. Que en la tabla MD analizaramos el ratio de ventas respecto de compras.
        - 1000 / 400

Modelado multidimensional

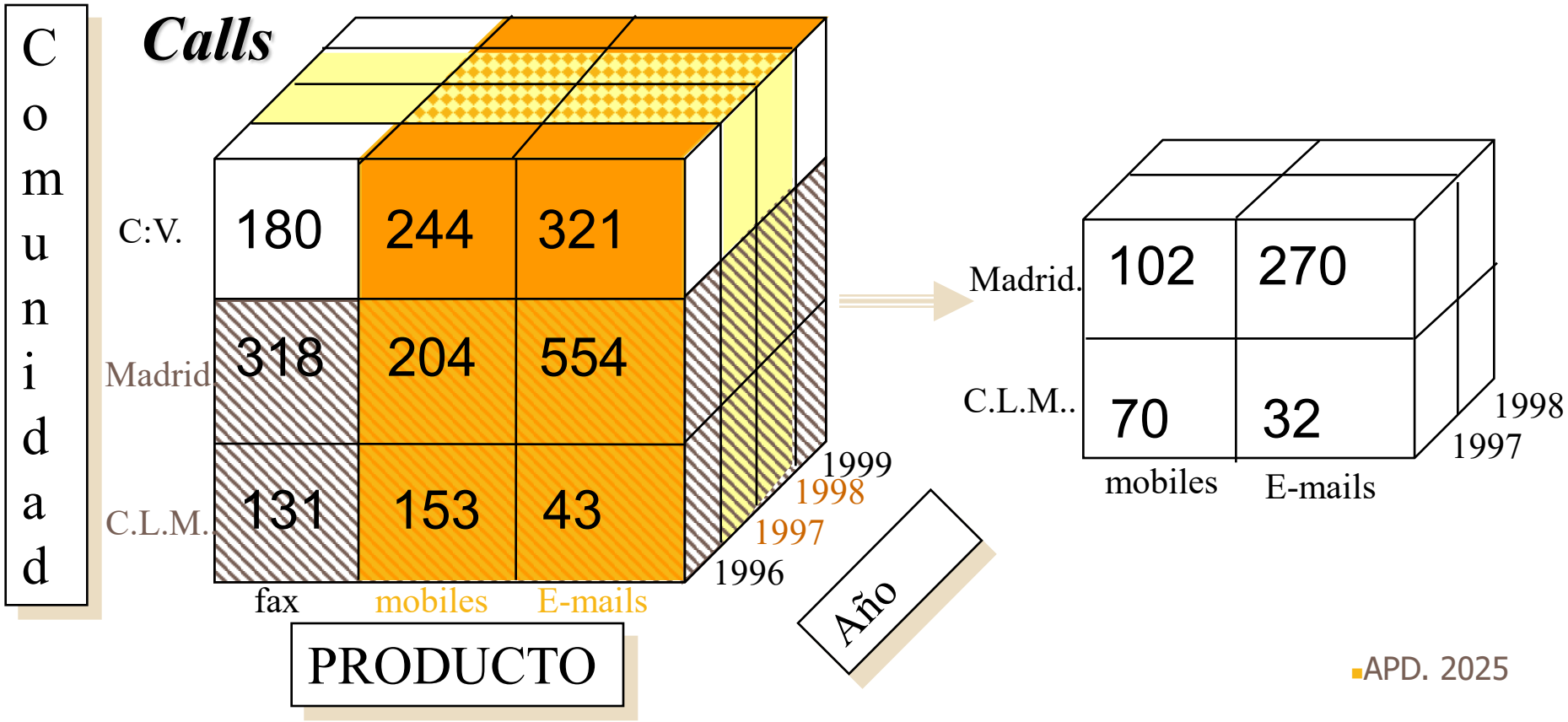
Parte dinámica.

- Operaciones de consulta (OLAP)
  - Slice-dice
    - Definir restricciones sobre niveles de jerarquías
      - Ej. Analizar datos donde el año sea 1999

Modelado multidimensional

Parte dinámica.

- Slice-dice (cont.)
  - Otro ejemplo



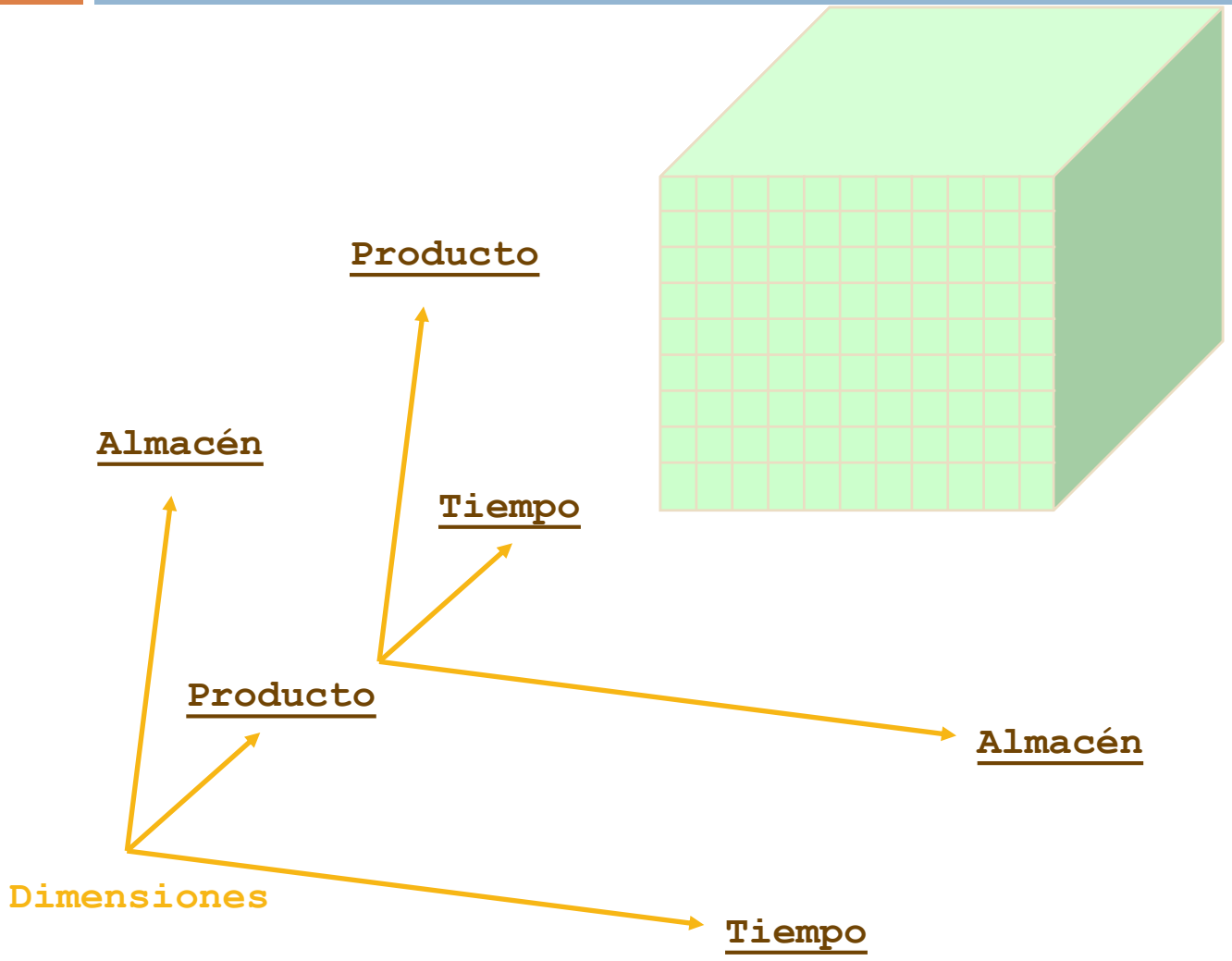
Modelado multidimensional

Parte dinámica.

- Operaciones de consulta (OLAP)
  - Pivoting
    - Reorientar la vista multidimensional de los datos, es decir, cambiar la distribución de filas/columnas
      - Algunos autores consideran también el intercambio de medidas y hechos como pivoting (kimball, 1996) (Inmon, 1996)

Modelado multidimensional

Parte dinámica.



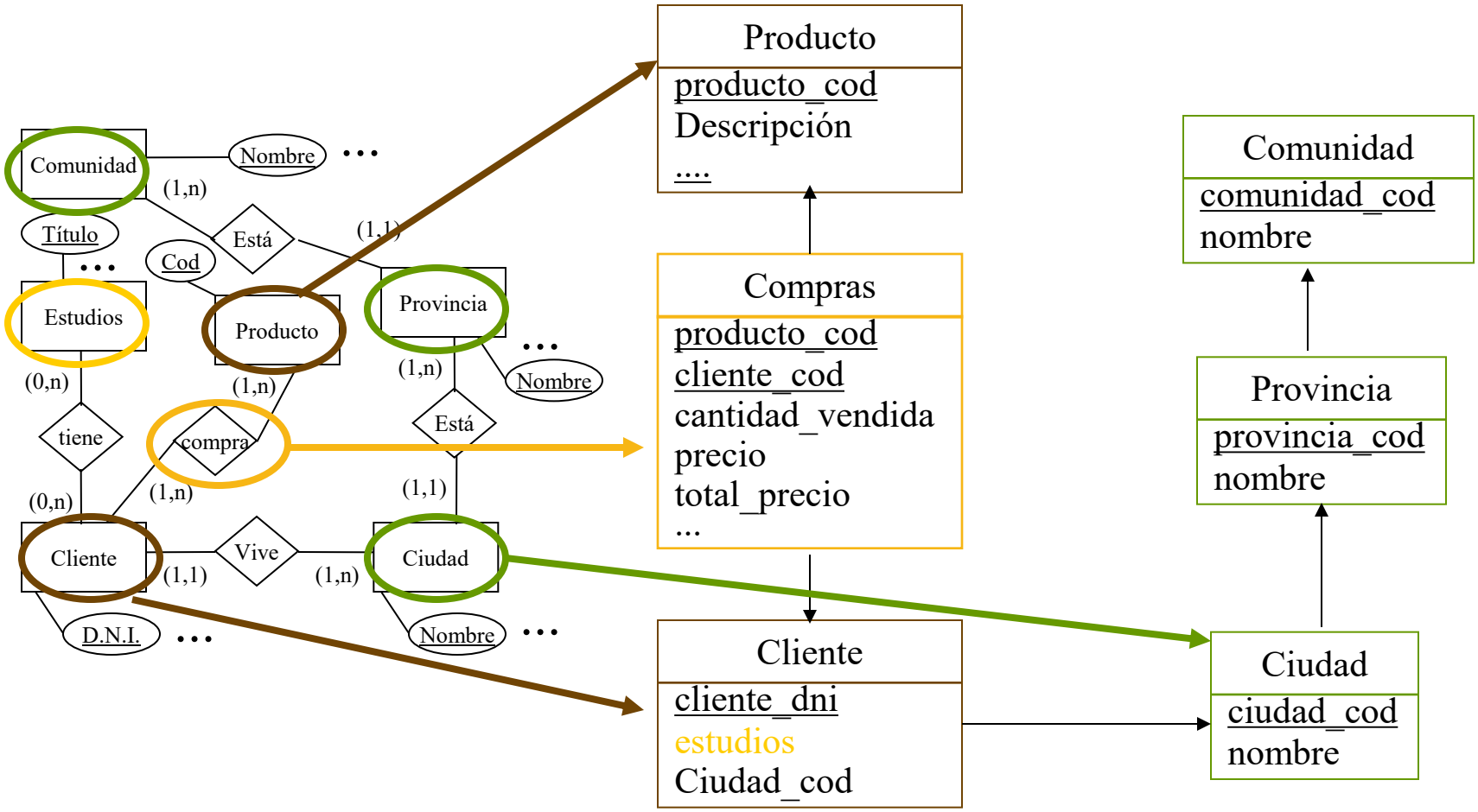
Diseño conceptual de almacenes de datos

Bibliografía

- Giovinnazo (2000). Object-Oriented Data Warehouse Design: Building a star schema
- Inmon (2005). Building the Data Warehouse (4ª ed.)
- Kimball & Ross (2013). The Data Warehouse Toolkit (4ª ed.)
- Thomsen (2000). OLAP solutions: Building Multidimensional Information Systems

Diseño conceptual de almacenes de datos

Apéndice: Guías de diseño del Modelo EER → Esq. Estrella



DISEÑO CONCEPTUAL DE ALMACENES DE DATOS

Tema 2

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Diseño lógico de almacenes de datos

Tema 3

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

- El modelado lógico multidimensional depende del tipo de servidor OLAP
  - MOLAP (Multidimensional OLAP)
    - Vectores o matrices multidimensionales
  - ROLAP (Relational OLAP)
    - Esquema estrella de R. Kimball
      - Tablas relacionales para representar hechos y dimensiones
      - Variantes esquema estrella
        - Constelaciones de hechos
        - Copos de nieve

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

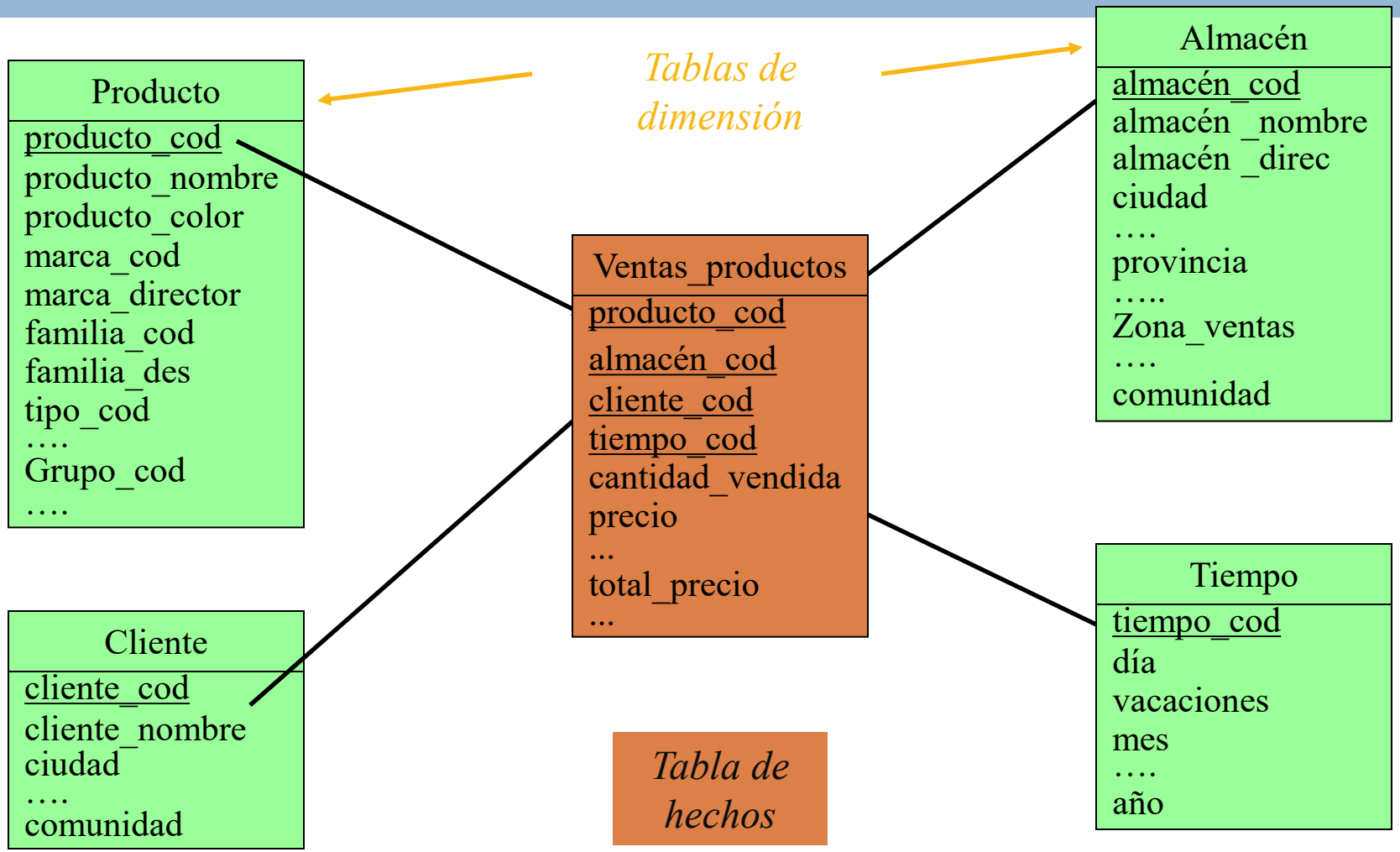
Modelado multidimensional

Esquema estrella

- Hechos → tablas de hechos
- Dimensiones → tablas de dimensiones
- Tabla de hechos es relación muchos a muchos con las tablas de dimensiones
  - Técnica utilizada → desnormalización
  - Clave primaria tabla de hechos compuesta por claves ajenas de tablas de dimensiones

Modelado multidimensional

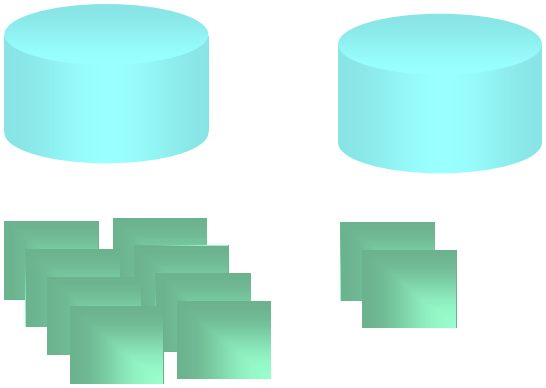
Esquema estrella



Modelado multidimensional

Esquema estrella. Granularidad.

- Granularidad
  - Afecta al Almacén de datos
    - Tamaño del repositorio
    - Grado de análisis
    - Flexibilidad
  - Nivel de detalle de los datos …
    - Transacciones individuales
    - Resúmenes diarios
    - Resúmenes mensuales
    - Resúmenes anuales
    - Cualquier otro período de tiempo

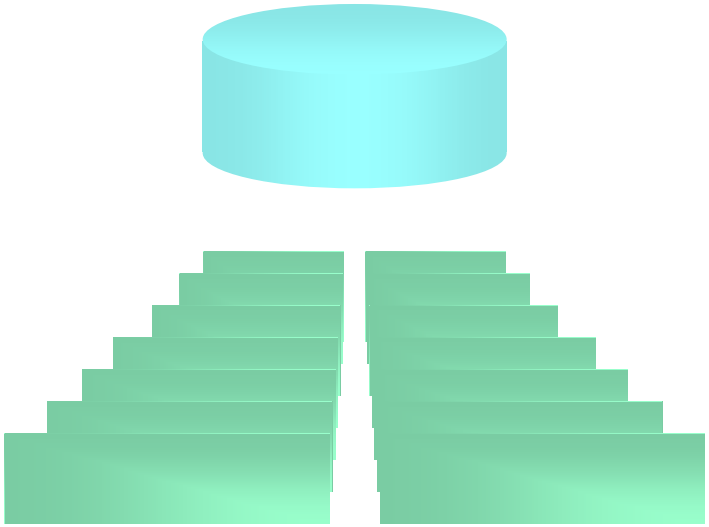


Modelado multidimensional

Esquema estrella. Granularidad.

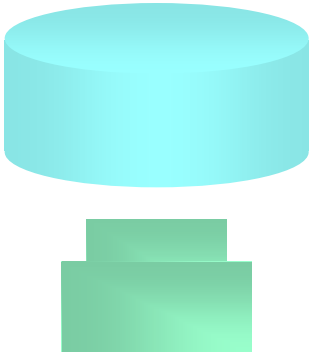
- Nivel alto (fine grain)

Detalles de las transacciones de los clientes del banco por mes



- Bajo nivel (coarse grain)

Resumen de las transacciones de los clientes del banco por mes



Define el nivel en función de las necesidades del negocio

Modelado multidimensional

Esquema estrella

- Tablas de dimensiones
  - Describen el contexto para analizar los hechos
    - Datos "textuales" (Alfanuméricos)
    - Datos desnormalizados → redundancia
  - Cada fila contiene su clave primaria y los atributos descriptores de todos los niveles de jerarquía
  - Tablas más pequeñas que las tablas de hechos

Modelado multidimensional

Esquema estrella

- *Tabla de dimensión producto*

Producto
producto_cod
producto_nombre
producto_color
marca_cod
marca_director
familia_cod
familia_des
tipo_cod
….
Grupo_cod
….

Producto_cod	Producto nombre ….	Family_Desc
1	Puleva Milk ½ L.	Productos lácteos
2	Puleva Milk 1L .	Productos lácteos
3	Yogourt Pascual	Productos lácteos
4	Mr. Proper	Productos limpieza
5	Ajax	Productos limpieza.
.	.	.
.	.	.

Modelado multidimensional

Esquema estrella

SQL*Plus Worksheet

Archivo Editar Hoja de trabajo Ayuda

ORACLE

```
--Connect jtrujillo/****@ibm

select * from almacen;

--set lin 800;

--desc almacen;
```

REG_ID	REG_NOMBRE	ALM_DIRECTOR	ALM_TIPO_EDIFICIO	CIU
1	Eastern	Jones	Modern	
2	Mid West	Smith	Original	
3	South East	Davis	Compact	
4	Pacific	Johnson	Modern	
4	Pacific	Green	Original	
1	Eastern	Brown	Compact	
1	Eastern	White	Modern	
5	South West	Williams	Original	
4	Pacific	Stuber	Compact	
5	South West	Merz	Modern	
2	Mid West	Erickson	Original	
2	Mid West	Kalman	Compact	
2	Mid West	Inmon	Modern	
6	Mountain	Strehlo	Original	
6	Mountain	Ollom	Compact	
2	Mid West	Mantle	Modern	
7	Mid Atlantic	Mays	Original	
3	South East	Maris	Compact	
1	Eastern	Ruth	Modern	
2	Mid West	Cobb	Original	
1	Eastern			

Modelado multidimensional

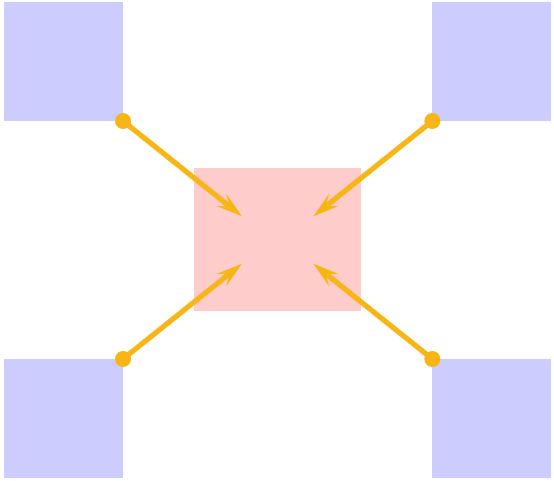
Esquema estrella

- Tablas de hechos
  - Actividades básicas de empresa
  - Cada fila se compone de:
    - Clave primaria (compuesta por claves ajenas de las dimensiones)
    - Medidas → Datos numéricos
  - Generalmente relación m-n con dimensiones y, m-1 en particular con cada dimensión

Modelado multidimensional

Esquema estrella

- *Tabla de hecho "ventas"*



CliKey	ProductoKey	AlmacénKey	Tiempo_key	Sale Amount
1	1	1	1	100€
1	2	1	2	120€
1	3	1	3	200€
.	.	.	.	

Modelado multidimensional

Esquema estrella

SQL*Plus Worksheet

Archivo Editar Hoja de trabajo Ayuda

ORACLE

```
--Connect jtrujillo/****@ibm

--set lin 2000;

select * from ventas;
```

60 filas seleccionadas.

ALM_ID	COSTES	DA_SMART_KEY	NUMERO_CLIENTES	PRM_ID	PRO_ID	UNIDADES_VENDIDAS	VENTAS
10	13,65	89	14	11	18	23	16,05
7	23,45	90	36	11	18	40	27,66
4	50,39	91	64	11	18	87	60,08
16	1,48	92	2	11	18	2	1,66
10	11,45	93	13	11	18	20	13,97
17	35,76	94	37	11	18	63	43,24
3	34,22	95	52	11	18	60	41,25
7	19,28	96	24	11	18	33	23,08
12	30,82	97	53	11	18	54	37,2
4	46,43	98	65	11	18	76	52,42
19	51,59	99	67	11	18	93	64,05
8	21,88	100	30	11	18	39	26,9
5	32,85	101	30	11	18	53	36,59
4	14,12	102	13	11	18	23	16,07
7	41,08	103	41	11	18	73	50,66
4	35,36	104	41	11	18	58	39,68
1	43,38	105	74	11	18	79	54,17

Modelado multidimensional

Esquema estrella

SQL*Plus Worksheet

Archivo Editar Hoja de trabajo Ayuda

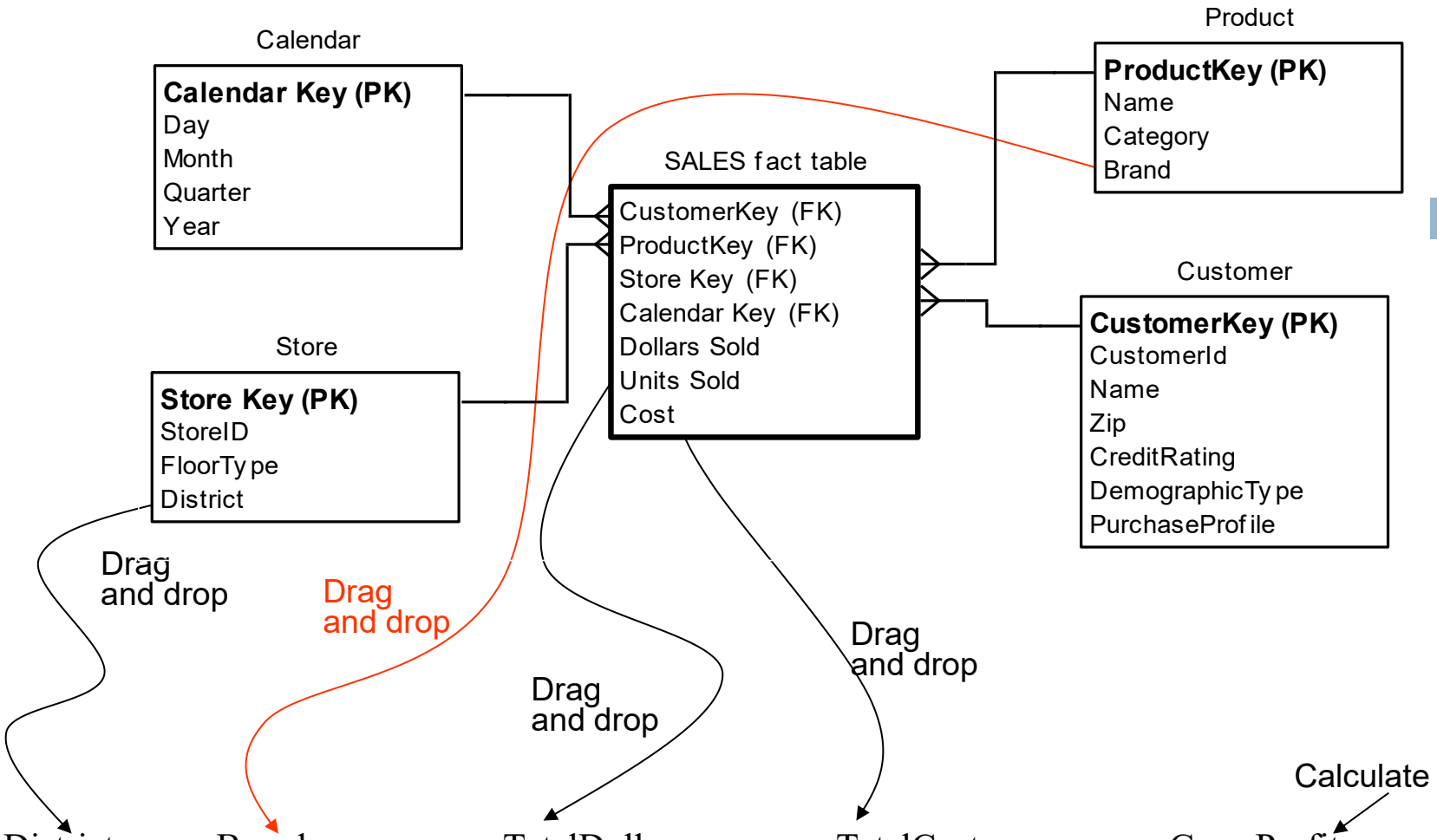
ORACLE

```
--Connect jtrujillo/****@ibm

--set lin 2000;

select * from ventas;
```

4	58,68	128	24	11	56	44	68,84
6	118,14	129	84	11	56	88	136,87
20	94,09	130	40	11	56	72	111,26
3	16,52	131	9	11	56	13	19,94
8	52,61	132	27	11	56	41	63,84
15	34,28	133	18	11	56	26	40,21
6	86,07	134	62	11	56	68	106,25
11	83,55	135	36	11	56	65	101,23
10	38,73	136	23	11	56	30	47
12	98,24	137	58	11	56	72	111,05
6	113,28	138	66	11	56	83	128,84
16	89,66	139	43	11	56	70	109,33
13	19,92	140	8	11	56	15	22,78
11	117,51	141	72	11	56	86	133,07
6	11,23	142	4	11	56	8	13,14
16	,43	143	0	11	56	0	,51
12	26,15	144	17	11	56	20	31,55
20	6,51	145	4	11	56	5	7,71

11044 filas seleccionadas.

Inicio | Microsoft Pow... | d:\orant9i\bin... | SQL*Plus Wo... | 19:05



District	Brand	TotalDollars	TotalCost	GrossProfit
Cook	Dorina	$2500	$2200	$300
Cook	Dado	$1000	$850	$150
Cook	Braco	$1000	$850	$150
DuPage	Dorina	$3000	$2600	$400
DuPage	Dado	$2000	$1850	$150
Dupage	Braco	$1000	$850	$150

Modelado multidimensional

Esquema estrella. Vista relacional.

SQL Statement:

```
select      District, Brand
            sum(DollarsSold) TotalDollars,
            sum(Cost) TotalCost,
            sum(DollarsSold) - sum(Cost) GrossProfit,
from        Sales, Store, product
where       Sales.store_key = Store.store_key and
            Sales.product_key = Product.product_key
group by    District, Brand
```

RESULT

District	Brand	TotalDollars	TotalCost	GrossProfit
Cook	Dorina	$2500	$2200	$300
Cook	Dado	$1000	$850	$150
Cook	Braco	$1000	$850	$150
DuPage	Dorina	$3000	$2600	$400
DuPage	Dado	$2000	$1850	$150
Dupage	Braco	$1000	$850	$150

Modelado multidimensional

Esquema estrella. Vista relacional.

Productos (dimensión en el esquema estrella)

Almacenes (loc. geog) (dimensión)

Cantidad

Producto	Ciudad	Día	Mes	Año	Unidades
Milk	Toronto	15	January	1997	1200
IntBread	Kingston	22	January	1997	1050
Juice	Ottawa	25	January	1997	1350
IntBread	Ottawa	28	January	1997	1500

Modelado multidimensional

Esquema estrella. Vista Multidimensional.

Corte en la tabla de dimensión: ***Enero***

Datos derivados

Almacén (dimensión)

Lugar

Producto	Toronto	Ottawa	Kingsto	Lake Reg.	Other Regs.	Total
Juice	0	1350	0	0	1350	1350
Milk	1200	0	0	1200	0	1200
IntBread	0	1500	1050	1050	1500	2550
Cheese	1600	0	0	1600	0	1600
Dairy	2800	0	0	2800	0	2800
Breads	0	1500	1050	1050	1500	2550
Beverages	0	1350	0	0	1350	1350

Producto (dimensión)

Cantidad y datos derivados

Corte en el Cubo Multidimensional

Modelado multidimensional

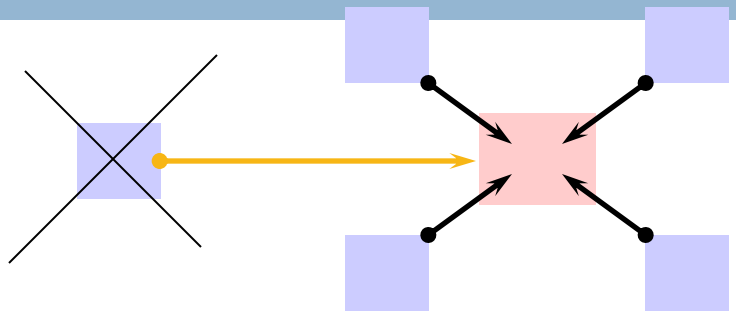
Esquema estrella. Particularidades.

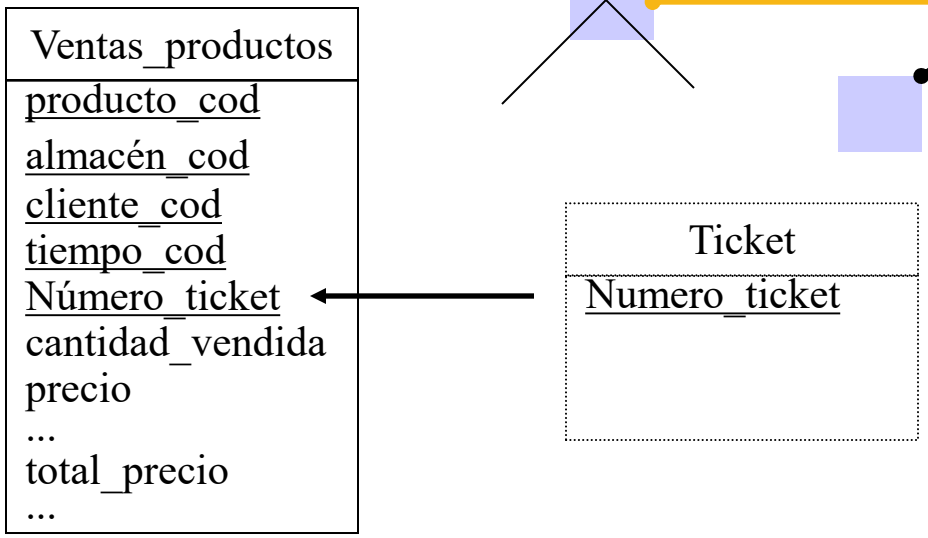
- Dimensiones degeneradas
  - Las tablas de hechos contienen información (atributo) sobre una dimensión que no existe físicamente
    - Este atributo puede/suele formar parte de la clave primaria de la tabla de hechos
      - Suelen referirse a dimensiones físicas en mundo real
        - Ej. Número de ticket, cod_factura, cod_albarán

Modelado multidimensional

Esquema estrella

- Dimensiones degeneradas





- Normalmente indica que hay relación muchos a muchos en particular con una dimensión
- A veces a estos esquemas se les denomina Multi Star Schema

Modelado multidimensional

Esquema estrella. Particularidades.

- Tablas de hechos que no son hechos (*Factless fact tables*)
  - No contienen medidas o atributos de análisis
  - Recogen la ocurrencia de un evento

Impartir_clase
Profesor_dni Asignatura_cod Alumno_dni ...

Modelado multidimensional

Esquema estrella. Particularidades

- Claves primarias de las tablas de dimensiones
  - (i) Autogeneradas
    - Rendimiento
    - Incluso en la dimensión Tiempo
    - Más fáciles de manejar para procesos ETL
  - (ii) Con significado semántico

Modelado multidimensional

Esquema estrella. Particularidades

- Otra vez !!!! → Razones para desnormalizar
  - Rendimiento
    - Eliminar el número de uniones entre tablas
      - Elevado número de filas
      - Unión es la operación más cara
  - Intuitivo
    - Más fácil para consultar por parte del usuario no experto

Modelado multidimensional

Esquema estrella. Particularidades

- Dimensiones que cambian lentamente
  - Las instancias de dimensiones se pueden considerar fijas a lo largo del tiempo
  - Sin embargo, a veces deseamos registrar cambios que puedan suceder en algunas dimensiones para seguir analizando datos históricos
    - Normalmente cambios en descripción de productos o de clientes
  - *Kimball* propone tres soluciones básicas y dos híbridas

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

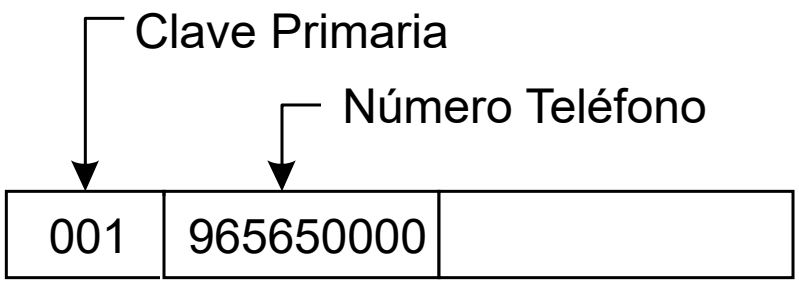
- Sol. 1: Sobreescribir el valor antiguo en la fila correspondiente de la dimensión
  - La fila registra el último valor válido
    - Perdemos la habilidad de analizar "la historia"
  - Se suele utilizar cuando el valor antiguo deja de tener significado o cuando se corrigen errores
  - Aproximación simple
  - Ejemplos: descripción productos, nombres nuevos, etc.

Modelado multidimensional

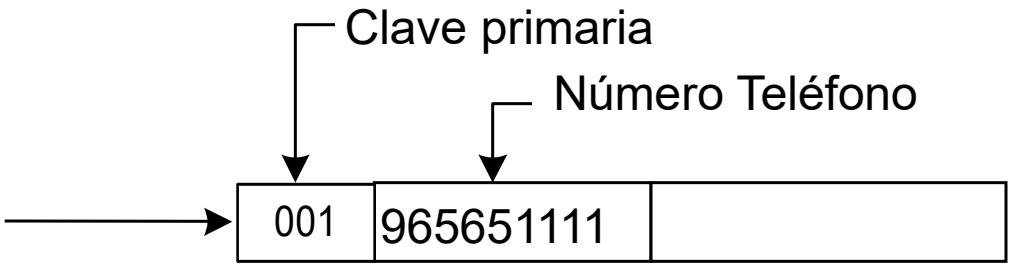
Esquema estrella. Dimensiones que cambian lentamente

- Sol. 1: Sobreescribir el valor antiguo en la fila correspondiente de la dimensión

Estado actual



Después de cambios



Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 2: Añadir una fila nueva utilizando un nuevo valor de clave autogenerada
  - Permite registrar el cambio del valor de un atributo en una dimensión
  - Representa un cambio físico real en la dimensión
  - Imprescindible utilizar claves autogeneradas
  - Añadir un campo que identifique al valor actual
    - Añadir un campo que sea un "flag" al *valor actual*

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 2: Añadir una fila nueva utilizando un nuevo valor de clave autogenerada…
  - Se puede utilizar dos atributos añadidos para registrar la fecha de comienzo y fin de validez de cada instancia
  - Inconvenientes
    - Necesita almacenamiento extra para los atributos extras
    - Realizar test para excluir las versiones antiguas
  - Ej. Dirección, Región…

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 2: Añadir una fila nueva utilizando un nuevo valor de clave autogenerada…
  - Segmenta la historia
    - Particiona la tabla de hechos
      - Dos "productos" distintos con claves generadas distintas estarán registrados en las ventas
  - Queries…
    - Definir restricciones sobre atributo que ha cambiado el valor
    - Si queremos mismo producto, actuaremos sobre clave primaria original

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 2: Añadir una fila nueva utilizando un nuevo valor de clave autogenerada…
  - Suele ser la más utilizada para las verdaderamente *Dimensiones que cambian lentamente*

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 2: Añadir una fila nueva utilizando un nuevo valor de clave autogenerada…

Estado actual

Clave primaria

CP

002	19104	

Después de cambios

Clave primaria

CP

003	19103	
002	19104	

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. 3: Añadir atributos para registrar el valor nuevo y el inmediatamente anterior
  - Suele utilizarse en cambios mínimos como por ejemplo redefinición de *descripción de producto*
  - Utilizado cuando deseamos seguir la historia de tanto los valores antiguos como nuevos
  - Se puede / suele añadir un atributo de *fecha_efectiva*
  - Poco recomendada / utilizada

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

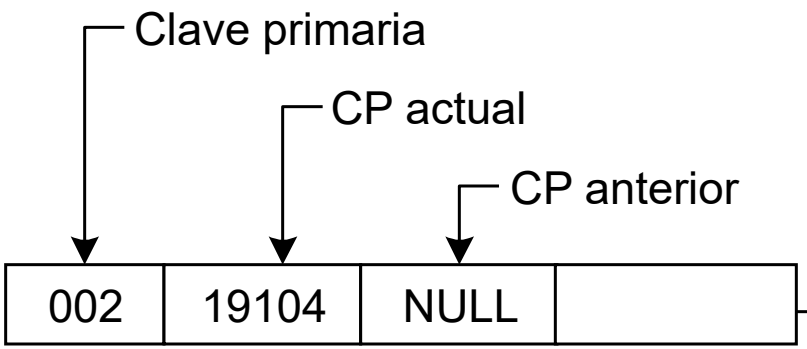
- Sol. 3: Añadir atributos para registrar el valor nuevo y el inmediatamente anterior
  - Inconveniente:
    - Sólo permite analizar los datos de la tabla de hechos o bien por el valor antiguo o bien por el nuevo
      - Ya que se utiliza la misma clave autogenerada

Modelado multidimensional

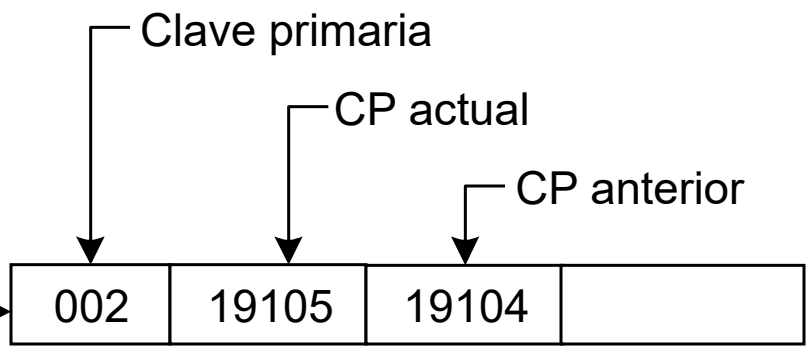
Esquema estrella. Dimensiones que cambian lentamente

- Sol. 3: Añadir atributos para registrar el valor nuevo y el inmediatamente anterior

Estado actual



Después de cambios



Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol híbridas son más flexibles aunque proporcionan mayor complejidad de manejo
  - Muy seguro de ser adecuado para el negocio y los posibles requerimientos que se pueden lanzar
- Sol. Híbrida 1: cambios predecibles con versiones múltiples que se superponen
  - Existen cambios predecibles y regulares
  - Usar la aproximación de tipo 3 generalizada

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. Híbrida 1: cambios predecibles con versiones múltiples que se superponen

Sales Rep Dimension
Sales rep key Sales rep name Sales Rep address Current district District 2001 District 2000 District 1999 ...

- Conocen los años de análisis de ventas
- Se puede consultar el estado del distrito en cada año

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Sol. Híbrida 2: cambios no predecibles con versiones simples que se superponen
  - Existen cambios irregulares e impredecibles
  - Necesidad de preservar los valores históricos
  - Ninguna de las soluciones estándares anteriores contempla estas dos características

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

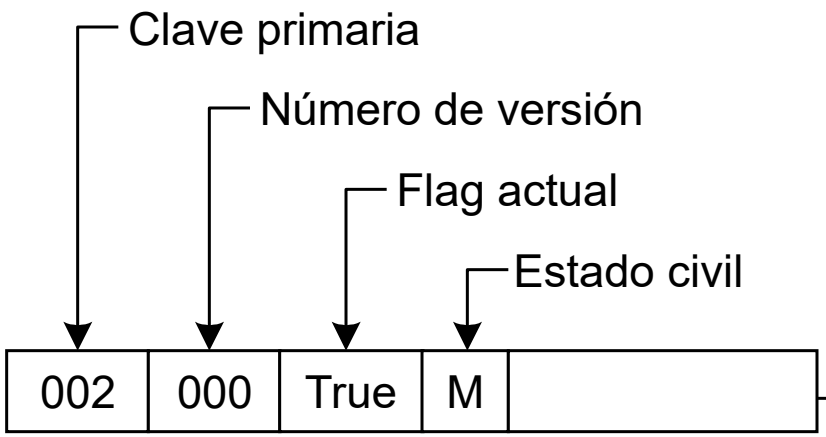
- Sol. Híbrida 2: cambios no predecibles con versiones simples que se superponen
  - Utilizar los tipo 1 + tipo 2 + tipo 3
    - Añadir una nueva fila para capturar los cambios
    - Añadir una nueva columna para controlar los valores actuales
    - Utilizar la aproximación de tipo 1 para cambios de valores muy recientes donde la historia no interesa
    - Más información: cap. 4 (*R. Kimball*), pag. 104

Modelado multidimensional

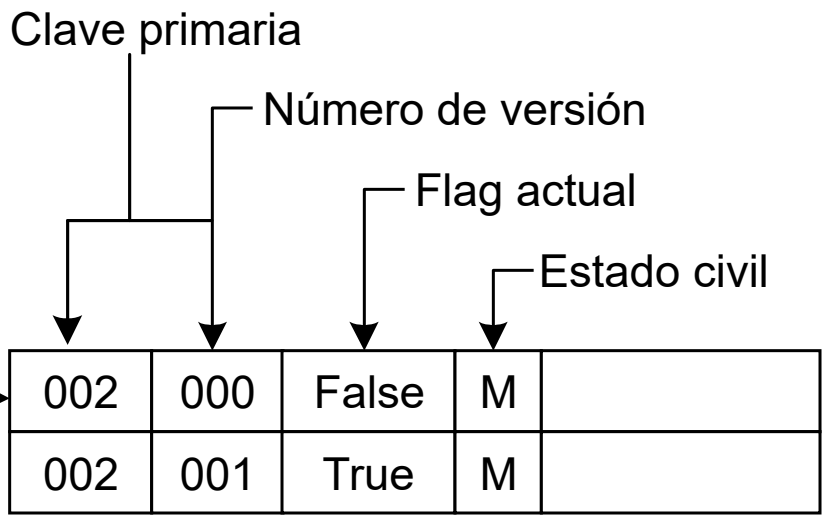
Esquema estrella. Dimensiones que cambian lentamente

- Otros métodos
- Método 4: Registrando versiones con *flag's*

Estado actual



Después de cambios



Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Método 5: Utilizando la clave primaria de los sistemas OLTP con flags

Estado actual

Clave primaria
Cliente ID
CP
Actual
Fecha

002	111	19104	True	5/31/2000

Después de cambios

Clave primaria
Cliente ID
CP
Actual Flag
Fecha

202	111	19107	True	5/31/2001
002	111	19104	False	5/31/2000

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Método 6: Modelar eventos con flag's
  - Utiliza claves de OLTP. Prácticamente tipo 2

Estado actual

Clave primaria

Estado civil

Momento del evento

002	13:00 1/1/2000	M	

Después de cambios

Clave primaria

Estado civil

Flag actual

Momento del evento

202	15:00 5/31/2001	S	True
002	13:00 1/1/2000	M	False

Modelado multidimensional

Esquema estrella. Dimensiones que cambian lentamente

- Método 7: Modelado del estado con clave OLTP y flag

Estado actual

Clave primaria	Fecha comienzo	Fecha final	Estado civil
002	09:00 1/1/1998	NULL	S

Después de cambios

Clave primaria	Cliente ID	Fecha comienzo	Fecha final	Estado civil	Flag actual
202	111	1/1/2000	NULL	M	True
002	111	1/1/1998	1/1/2000	S	False

Modelado multidimensional

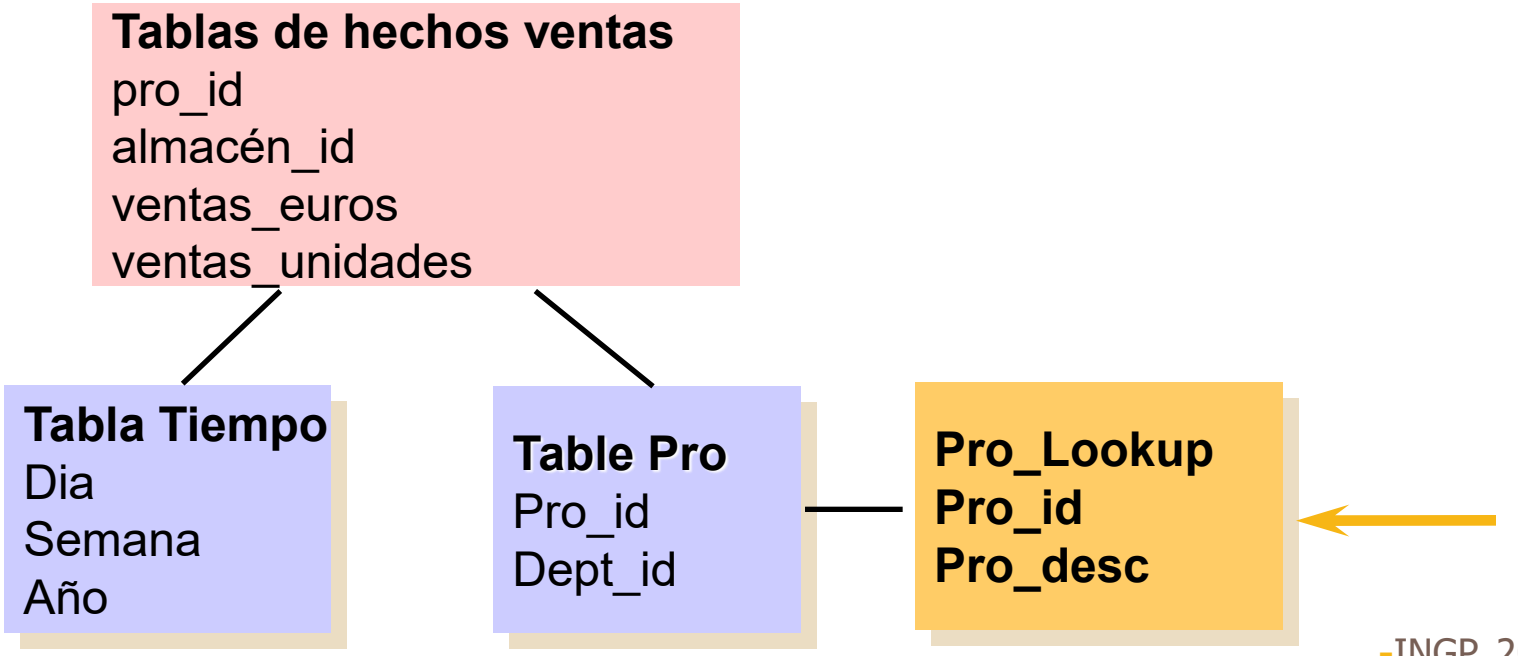
Esquema estrella

- Qué información podemos saber a partir del esquema estrella ....
  - ¿Cuál es la jerarquía de clasificación en las dimensiones?
  - ¿Es la jerarquía entre almacén y zona_Ventas estricta?
  - Hay demasiada información explícita → necesidad de documentar el diseño

Modelado multidimensional

Esquema estrella. Tablas y datos de referencia.

- Proporciona apoyo para la gestión de las dimensiones
- Se reduce el volumen del almacén de datos



Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

Esquema constelaciones de hechos

- Esquema de constelaciones de hechos (fact constellations)
  - Además de la tabla de hechos base, se define una tabla de hechos para cada nivel de agregado
  - A veces se les denomina *Multi Fact Tables* por contener más de una tabla de hechos
  - Acelera la consulta de datos agregados

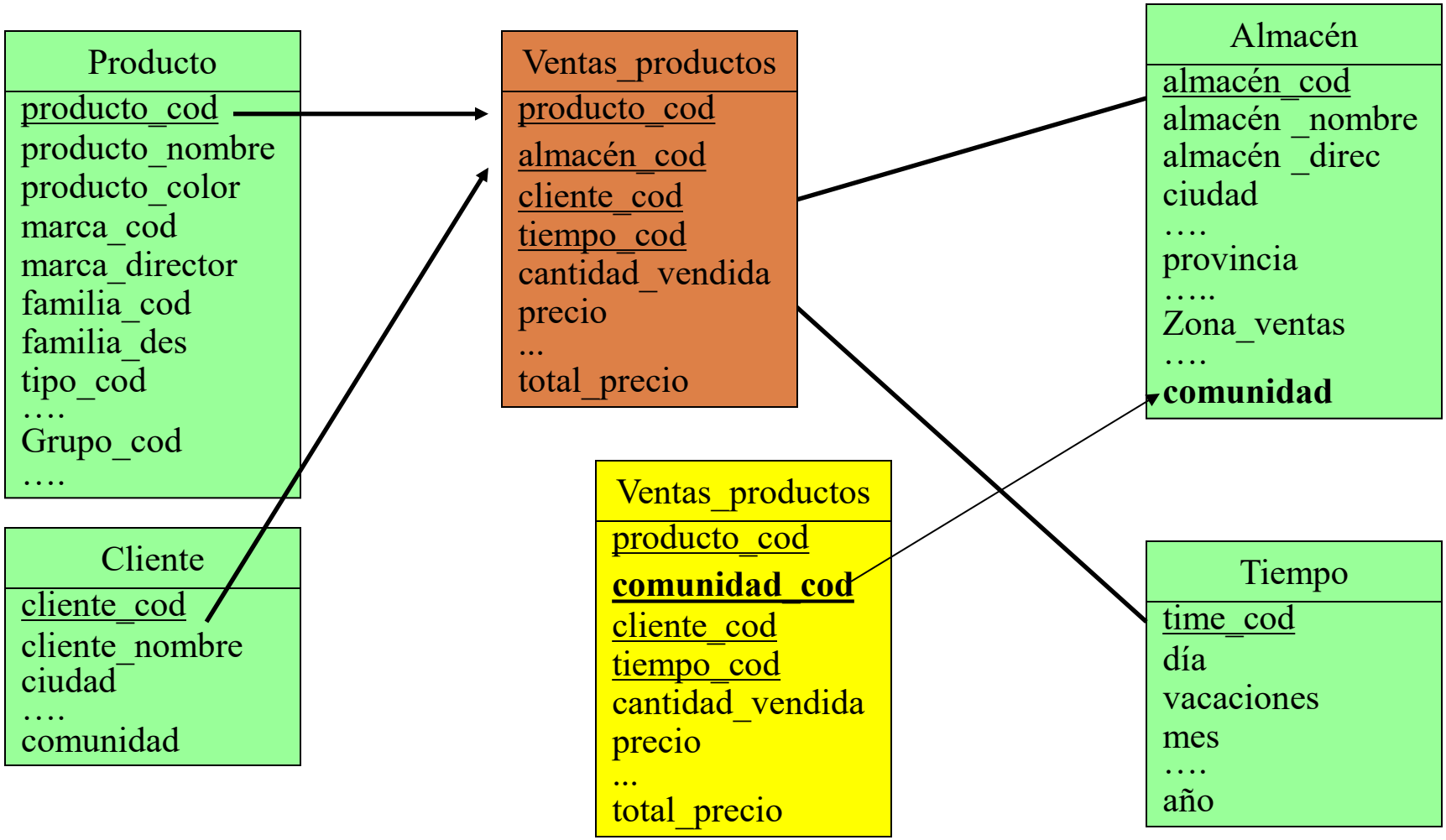
Modelado multidimensional

Esquema constelaciones de hechos

- Esquema de constelaciones de hechos (fact constellations)
  - Inconvenientes
    - Cada tabla de agregados se usa para calcular su nivel
      - Navegar por jerarquías requiere escanear (browse) distintas tablas
    - Aumenta el tamaño de los metadatos
    - Dificulta su gestión y mantenimiento ya que para cada carga nueva de datos se han de recalcular todos las tablas de hechos
    - Puede haber requerimientos que necesiten varias tablas

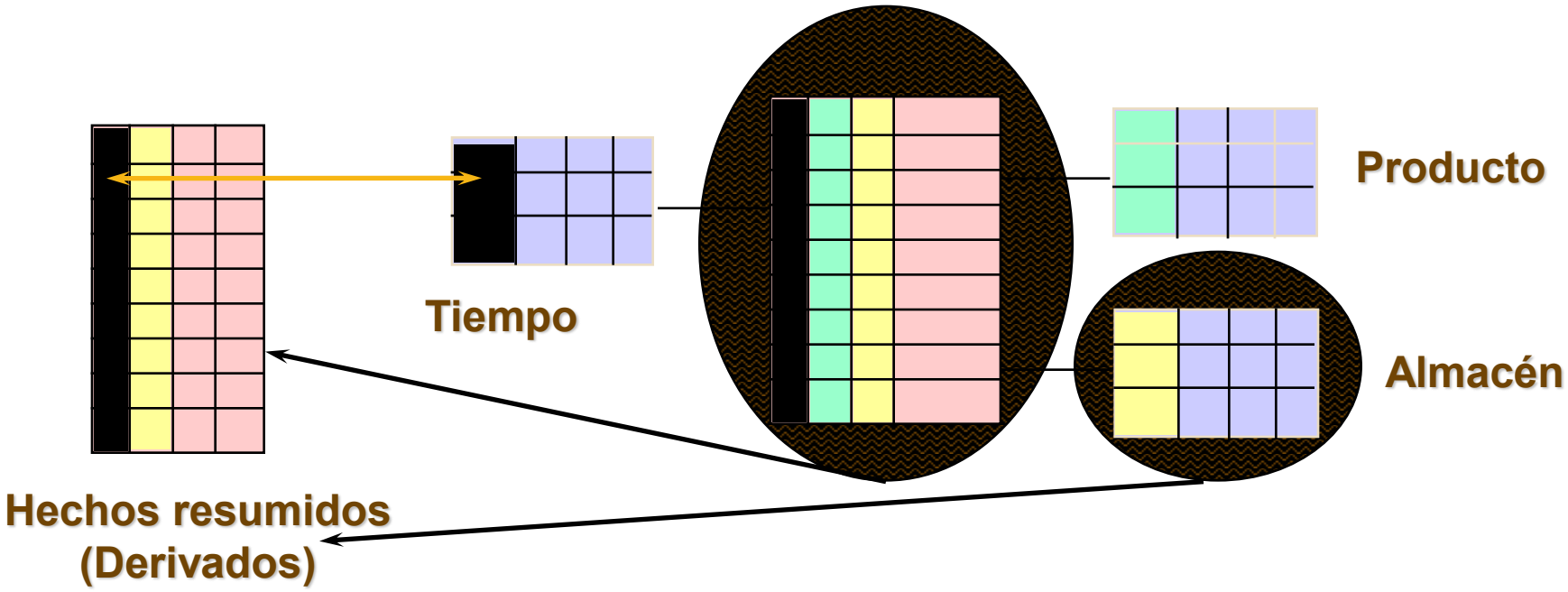
Modelado multidimensional

Esquema constelaciones de hechos



Modelado multidimensional

Esquema constelaciones de hechos



Modelado multidimensional

Esquema constelaciones de hechos

- Acceso rápido a datos precalculados
- Reduce el uso de
  - I/O
  - CPU
  - Memoria
- Diferencia con fuentes…
  - Sistemas fuente de datos – poco resumidos
  - Resúmenes precalculados – muy resumidos
- Determinar los requerimientos pronto

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

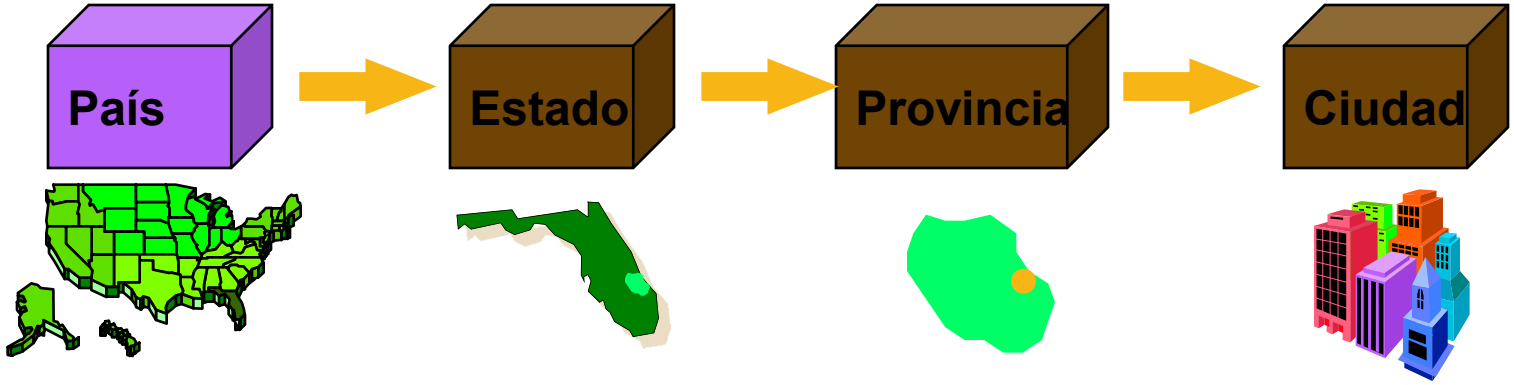
Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

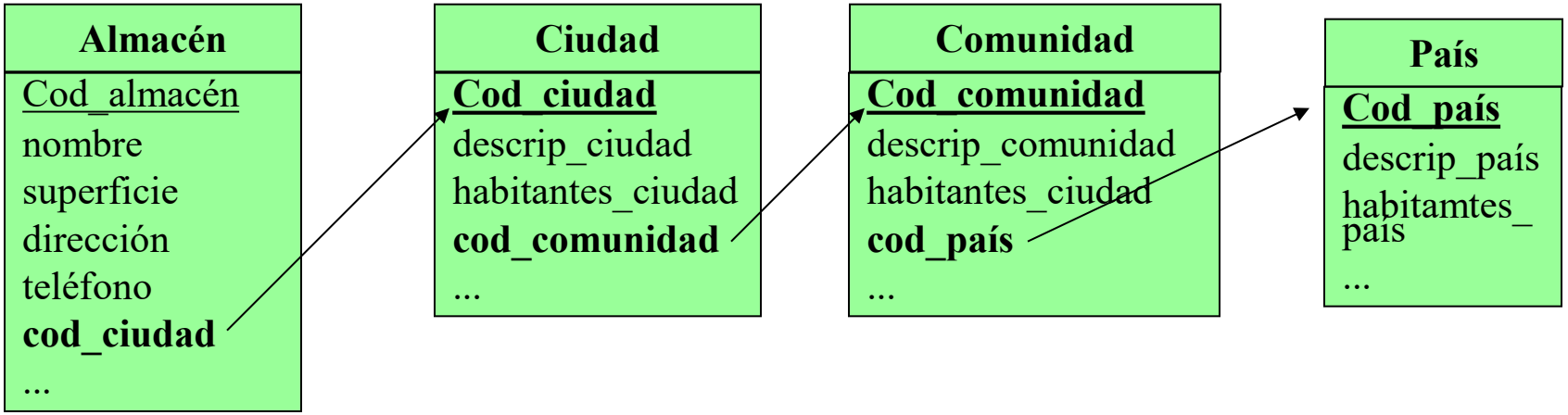
Esquema de copos de nieve

- Las dimensiones se normalizan con los niveles de jerarquía de dimensiones
  - Tabla dimensión → valores del mínimo nivel de jeraquía



Modelado multidimensional

Esquema de copos de nieve



Modelado multidimensional

Esquema de copos de nieve

- Ventajas
  - Podría salvar espacio en disco, pero no demasiado
  - Consultas (browsing) de atributos simples es rápido
  - Mejora el rendimiento cuando mayoría de requisitos solicitan niveles de agregación superiores
    - Disminuye el tamaño de tablas a escanear

Modelado multidimensional

Esquema de copos de nieve

- Inconvenientes
  - Aumenta el número de tablas → aumenta el número de uniones (join)
    - Algunos requisitos pueden demorarse en exceso
  - Consultar (browsing) atributos de más de una tabla es más lento
  - Demasiado complejo para que los usuarios finales definan sus propias consultas ad-hoc a partir de ellos

Modelado multidimensional

Esquema de copos de nieve

- Inconvenientes
  - Requiere una clave primaria más por cada nivel de jerarquía normalizado
  - Aumenta la complejidad de diseño y mantenimiento
  - No soportado por todas las herramientas del mercado

Modelado multidimensional

Esquema de copos de nieve

- Recomendaciones
  - Normalmente no se recomienda
    - Realmente cuando el espacio en disco es un problema
  - Normalmente se recomienda normalizar una o dos de las dimensiones más grandes
    - Existen un gran número de filas
  - Suele aplicarse cuando muchos atributos caracterizan a los niveles más altos de las jerarquías

Modelado multidimensional

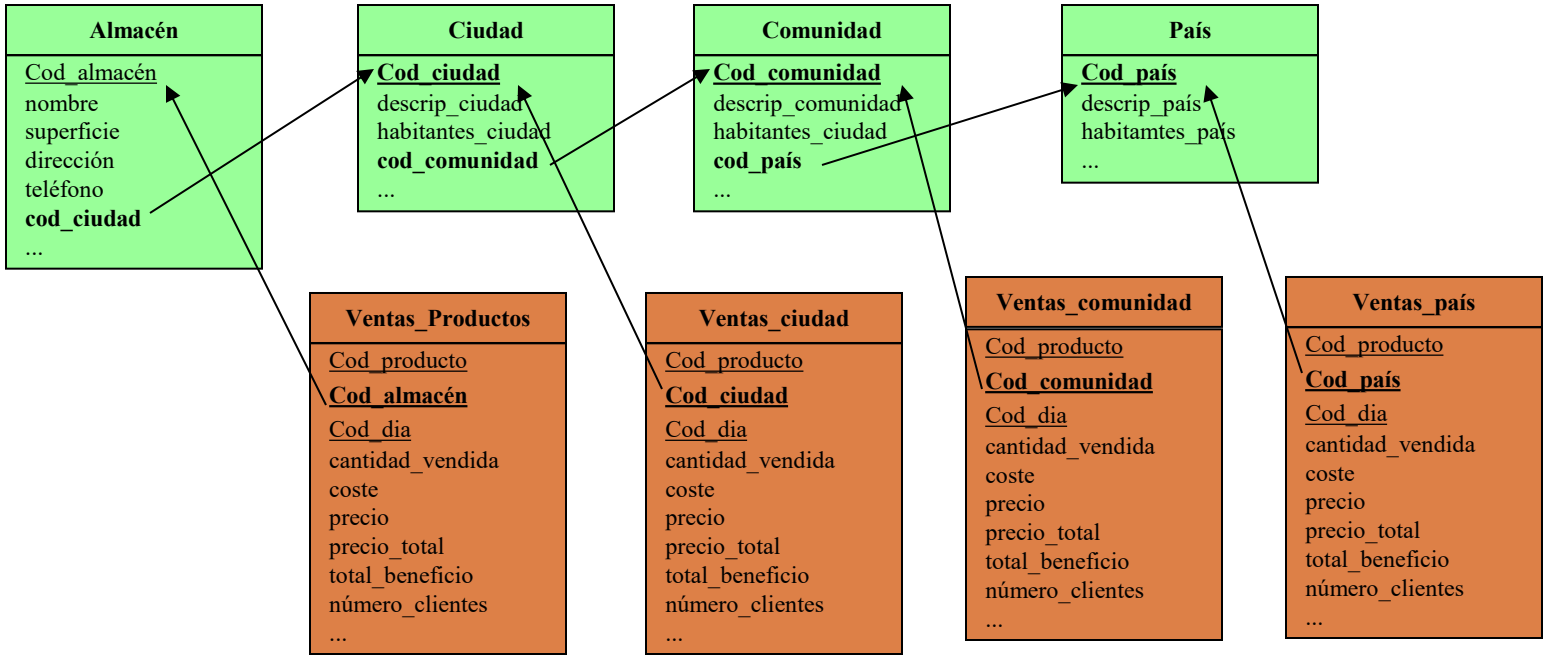
Esquema de copos de nieve

- Recomendaciones
  - Cuando se pueden utilizar índices bitmap para los atributos de granularidad mínima
  - Utilizar sólo cuando las ventajas son muy explícitas:
    - Ahorro en disco significativo
    - Muchos atributos en los niveles más altos de jerarquías
  - Estadísticamente, el espacio en disco ahorrado utilizando *snowflake schemas* es del 1% del espacio total en disco

Modelado multidimensional

Esquema de copos de nieve y constelaciones de hechos

- Utilizar copos de nieve con constelaciones de hechos (agregados)



Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema estrella
  - Ventajas
    - Fácil de entender por los usuarios
    - Reduce número de uniones físicas
      - Respuestas rápidas para la mayoría de las consultas
      - Fácil de entender por usuarios finales
    - Metadatos sencillos
    - Soportado por la inmensa mayoría de aplicaciones

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema estrella
  - Inconvenientes
    - El aumento del tamaño de la tabla de hechos con datos agregados puede empeorar el rendimiento general
      - Por ello se recomienda tablas de hechos agregados al margen
    - Las dimensiones tienen un tamaño enorme
      - Alrededor de 50 atributos (*Kimball*)
    - La necesidad de manejar el atributo de nivel para datos agregados en algunos sistemas
    - Es poco robusto o susceptible a cambios
    - Más lento de construir

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema constelaciones de hechos
  - Ventajas
    - No se necesita el atributo de nivel en las tablas de dimensiones que lo utilizan explícitamente
    - Los datos agregados no se almacenan con los del nivel de detalle más bajo en las tablas de hechos
      - Aumenta el rendimiento al hacer browse de datos almacenados precalculados (sumados)

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema constelaciones de hechos
  - Inconvenientes
    - Un gran número de tablas de agregados
      - Complica el mantenimiento de los metadatos
      - Necesidad del usuario de conocer la existencia de dichas tablas
    - Algunos requisitos pueden necesitar consultar datos de varias tablas y mermar el rendimiento general del sistema

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema copos de nieve
  - Ventajas
    - No se necesita el atributo de nivel en los sistemas que lo gestionan explícitamente
    - Más flexible y adecuado para requerimientos
    - Carga de datos (ETL) más rápida y sencilla

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema copos de nieve
  - Ventajas
    - Directamente implementados por algunas jerarquías
      - No Todas
    - Mejora considerablemente el rendimiento cuando un gran número de requisitos solicita datos agregados o de niveles superiores de jerarquías
      - Los requerimientos escanean un reducido número de filas

Modelado multidimensional

Resumen/comparativa final del esquema estrella y sus variantes

- Esquema copos de nieve
  - Inconvenientes
    - Aumenta la complejidad de mantener los metadatos
      - Debido al aumento del número de tablas
    - Si no se dispone de la suficiente cantidad de tablas de agregados, el rendimiento general podría disminuir

Diseño lógico de DW

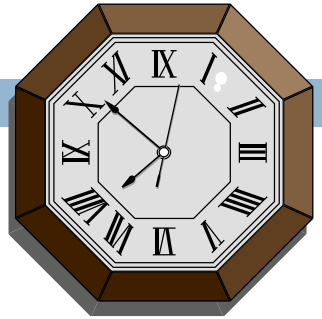
- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

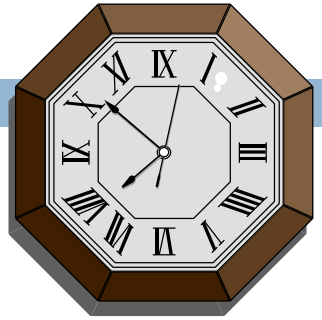
La dimensión tiempo



- La dimensión tiempo
  - Es obligatoria
  - Es única
  - Es potente
  - Necesita ser diseñada cuidadosamente
- Incluye períodos del tiempo de negocio
- Incluye fechas especiales

Modelado multidimensional

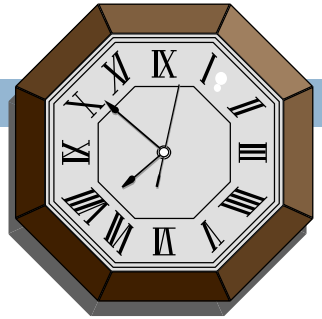
La dimensión tiempo



- Analizar y planificarla
- Considerar rango de fechas relevante
- Identificar la granularidad
- Utilizar las características de forma uniforme para navegar por la dimensión
- Modelar de forma flexible

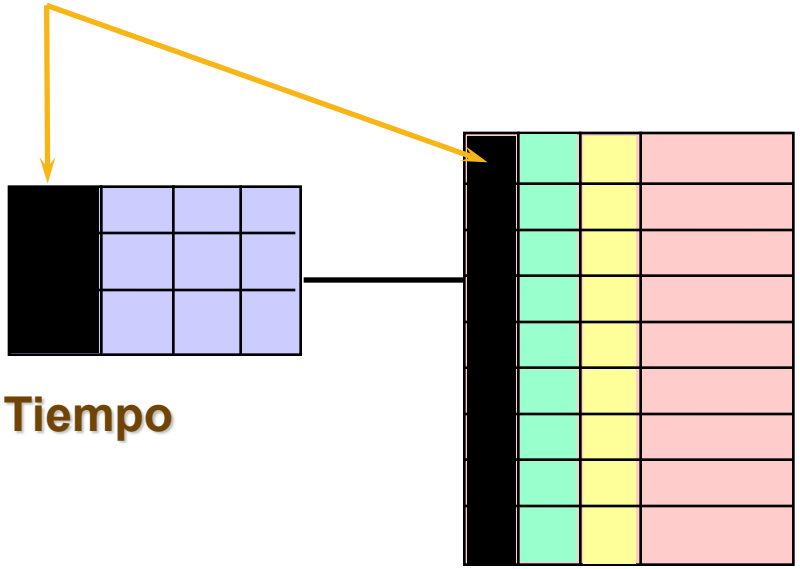
Modelado multidimensional

La dimensión tiempo



- Tabla de dimensión tiempo
  - Día laborable
  - Período fiscal
  - Principal evento
  - Mes
  - Vacaciones

Clave de la Tabla tiempo
Una columna simple: generl. auto generada

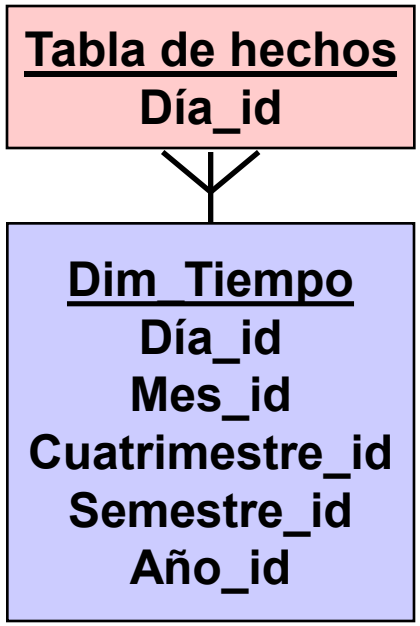


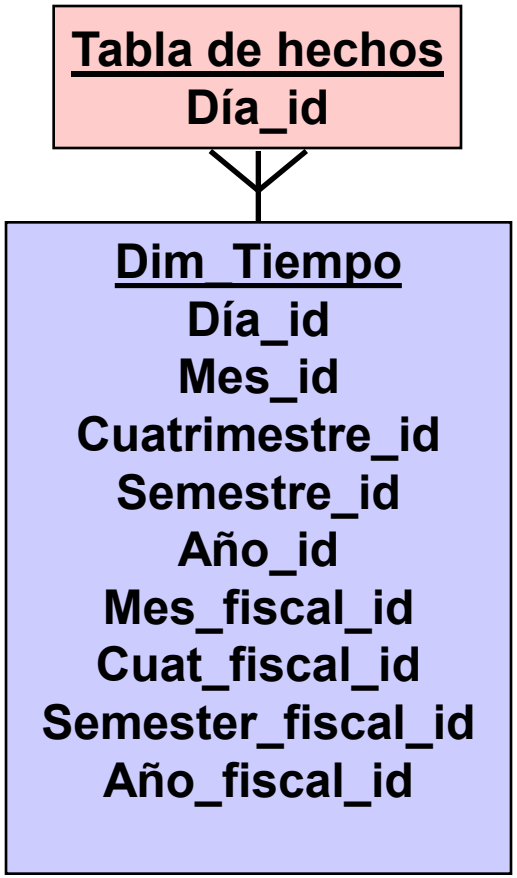
- Permite un análisis más flexible

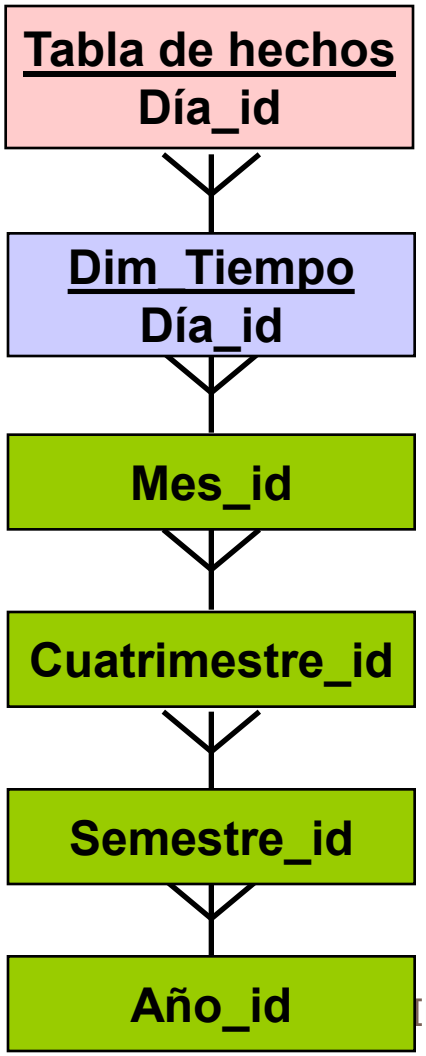
Modelado multidimensional

La dimensión tiempo

- Normalizada



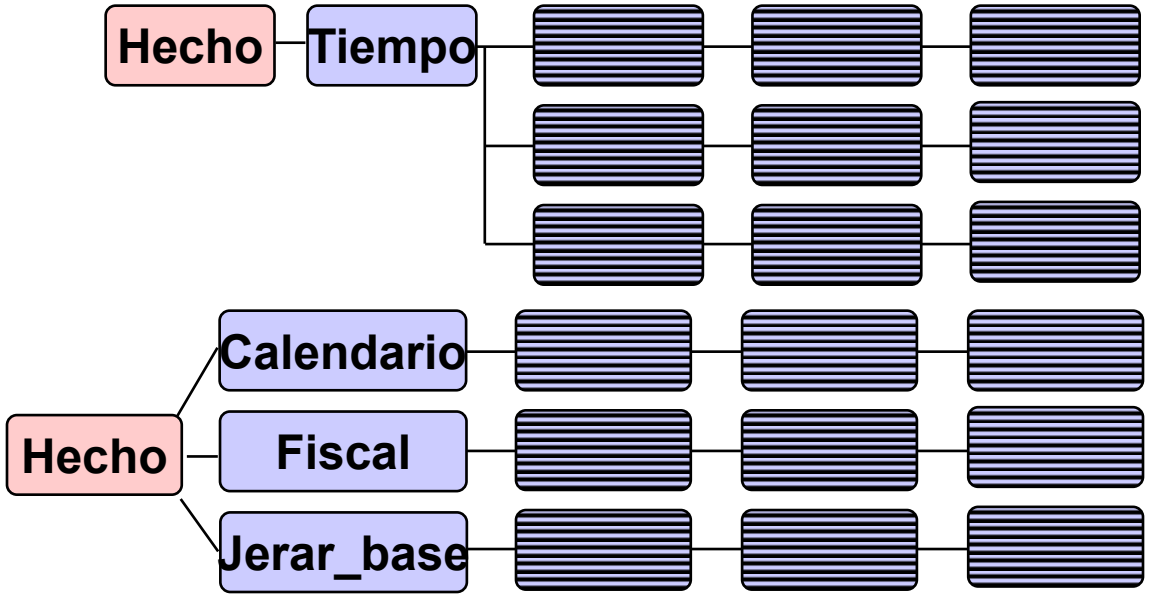




Modelado multidimensional

La dimensión tiempo

- Identificar requerimientos fecha
  - Normales
  - Fiscales
  - Jerarquías

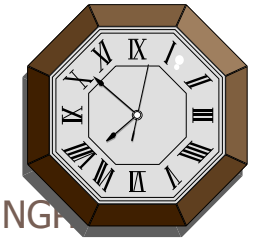


- Crear ...
  - Jerarquías múltiples
  - Dimensiones independientes
  - Vistas

Modelado multidimensional

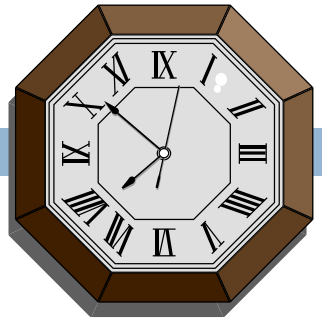
La dimensión tiempo

- Tipo de consultas sobre dimensión tiempo
  - Absolutas
    - Ej. Ventas desde julio 1999 hasta julio 2000
  - Relativas
    - Ej. Obtener ventas para el mes actual del año pasado



Modelado multidimensional

La dimensión tiempo



- Algunos sistemas de forma interna
  - Añaden un atributo para indicar período actual y el previo
    - Suele hacerse en el mínimo nivel de granularidad
    - Aunque hay que definir muy bien los periodos de comparación
  - Proporcionar códigos de tiempo secuencial que identifiquen a los elementos de las jerarquías

Modelado multidimensional

La dimensión tiempo

Cod Dia	Día	Cod Mes	Mes	Cod Semestre	Semestre	Cod Año	Año	Periodo actual	Agg level
1300	2	500	Sept	200	S3	50	1999		1
1301	3	500	Sept	200	S3	50	1999	**P**	1
…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…
1420	1	504	Enero	202	S1	51	2000		1
1421	2	504	Enero	202	S1	51	2000		1
…	…	…	…	…	…	…	…	…	…
1665	2	512	Sept	204	S3	51	2000		1
1666	3	512	Sept	204	S3	51	2000	**S**	1

- Tabla de dimensión tiempo en Informix Metacube

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

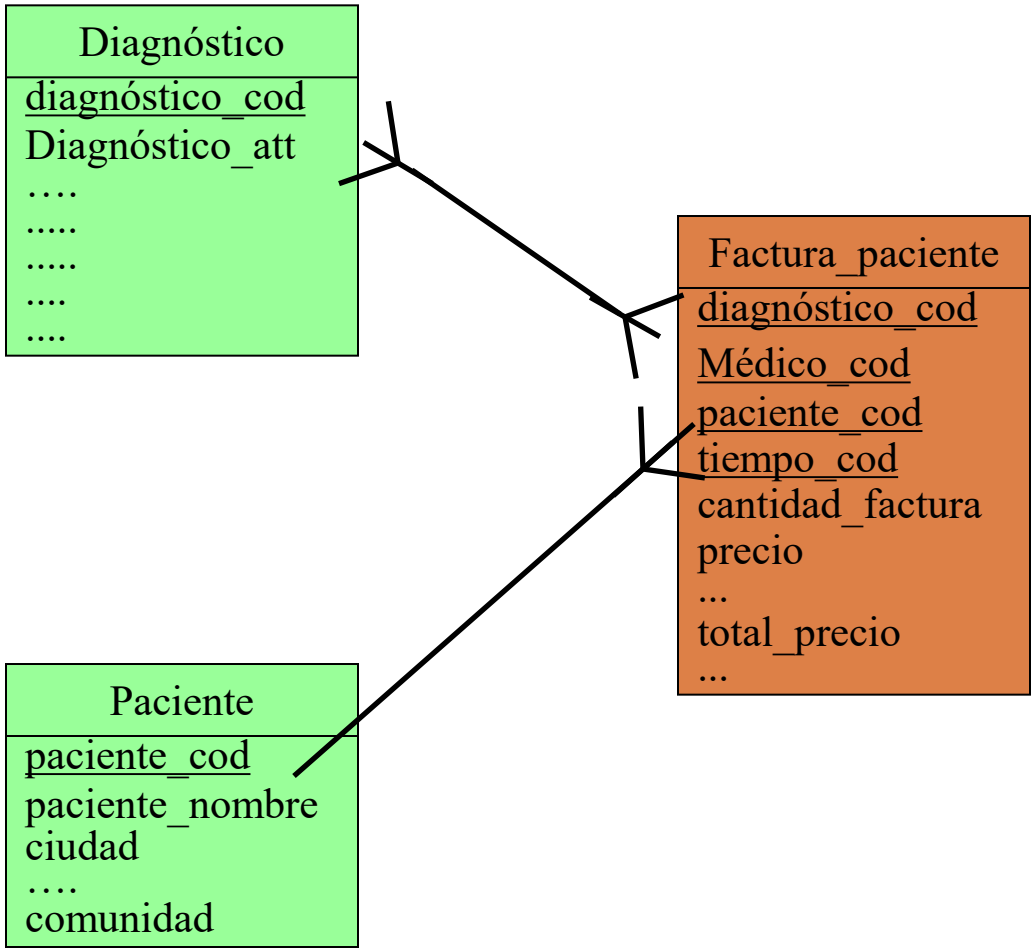
Modelado multidimensional

Otras consideraciones de diseño

- *Many-to-many dimensions*
  - Cuando entre la tabla de hechos y una tabla de dimensión existe una relación de muchos a muchos (m-n)
  - Ejemplo: Health Care System
    - Un paciente puede tener varios diagnósticos
    - ¿Qué hacer cuando un paciente tiene varios diagnósticos?

Modelado multidimensional

Otras consideraciones de diseño



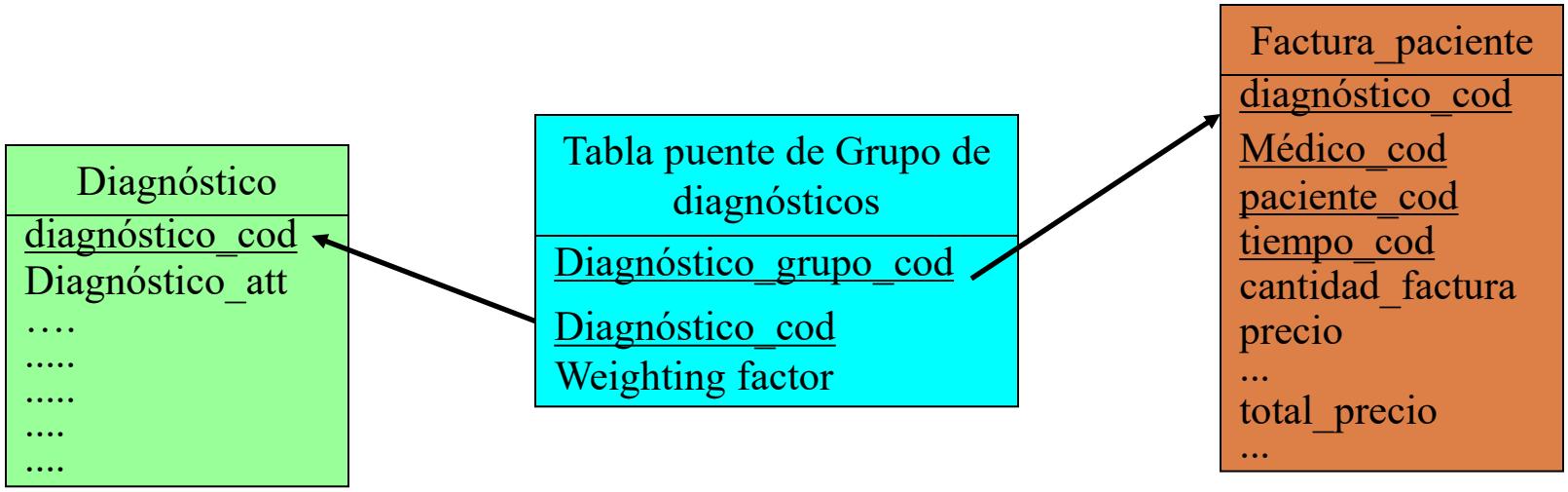
Modelado multidimensional

Otras consideraciones de diseño

- *Many-to-many dimensions*
  - Problemas
    - Sólo se podría visualizar uno de forma sencilla
    - Para visualizar todos, la generación del informe es más compleja
      - Necesidad de más uniones entre tablas
    - Se suele utilizar los denominados *weighting factors*
    - Kimball propone el uso de tablas puente (*bridge tables*)

Modelado multidimensional

Otras consideraciones de diseño



Modelado multidimensional

Otras consideraciones de diseño

- *The Data Warehouse Bus Arquitecture*
  - Arquitectura detallada de cómo construir todo el DW corporativo
  - Cada Data Mart debería estar basado en la mínima granularidad del proceso de negocio que contempla
  - Dos aproximaciones:
    - Crear una arquitectura que defina todo el DW
    - Supervisar la construcción de cada Data Mart en particular
  - Utilizar las dimensiones y hechos comunes (*conformed*)

Modelado multidimensional

Otras consideraciones de diseño

- Dimensiones comunes (*Conformed dimensions*)
  - Es una misma dimensión que existe en todos los DM
    - Ejplo. Tabla maestro de clientes o productos
      - Tablas que se mantienen de forma independiente a los DM
  - Cuando se definen dimensiones comunes
    - No obviarlas ni menospreciarlas
    - Definir el mínimo nivel de granularidad posible
    - Utilizar claves auto generadas (*surrogate keys*)
  - Este tipo de dimensiones puede ocupar el 80% del esfuerzo del desarrollo total

Modelado multidimensional

Otras consideraciones de diseño

- Dimensiones comunes (*Conformed dimensions*)
  - Ventajas
    - Una misma dimensión se puede utilizar contra varios hechos
    - Interfaces de usuario y datos son consistentes
    - Permite navegar entre Data Marts (DM)
- Hechos comunes (*Conformed facts*)
  - Medidas utilizadas en más de un DM
    - Ejemplos: beneficio, coste, precio etc.

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Diseño lógico de DW

- Modelado lógico multidimensional (MD)
- Modelado multidimensional relacional
  - Esquema estrella
  - Esquema constelaciones de hecho
  - Esquema copos de nieve
  - Resumen/comparativa final
  - La dimensión tiempo
  - Otras consideraciones de diseño
- El proceso de diseño de un DW

Modelado multidimensional

El proceso de modelado de un almacén de datos

- A partir de análisis de requisitos
  - Objeto de análisis → hechos
  - Contexto de análisis → dimensiones
    - Jerarquías, etc…
  - A veces no es intuitivo y se adoptarán decisiones sobre si considerar un elemento como dimensión o como hecho

Modelado multidimensional

El proceso de modelado de un DW: ¿¿¿ Hechos o dimensiones ????

Unidades vendidas – Llamadas en un mes - Color

Guardarlo en la tabla de dimensión si el atributo se percibe como un valor estático o constante:

- Descripción
- Localización
- Color
- Tamaño

Almacenar en la tabla de hechos si el atributo varía continuamente

- Balance
- Unidades vendidas
- Llamadas por mes

?

Coste producto

?

Modelado multidimensional

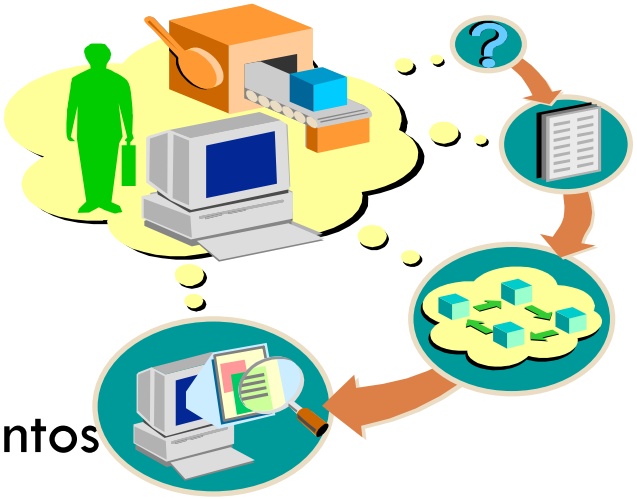
El proceso de modelado de un DW

- ¿¿¿ Diseño conceptual ???
  - Algunos diseñadores lo prefieren por la documentación
- Diseño lógico a partir del modelo corporativo/transac.
  - Extraer de la documentación o modelo → esquema estrella
    - Tablas de hechos
    - Tablas de dimensiones
      - Jerarquías de clasificación
      - Atributos
    - En función de los requisitos más comunes decidir
      - Normalizar algunas dimensiones → copos de nieve
      - Estudiar qué agregados definir → constelaciones de hechos

Modelado multidimensional

El proceso de modelado de un DW

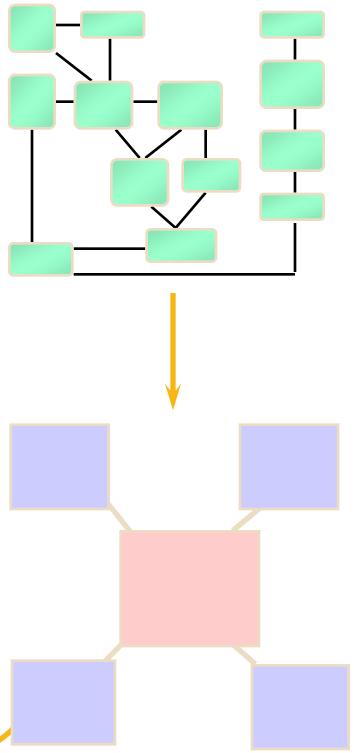
- Determinar los requerimientos de usuario
- Asistir a los usuarios en la comprensión de la tecnología
- Casi necesario…
  - Definir requerimientos analíticos
  - Objetivos de rendimiento
- Mantenerse informado con los cambios de requerimientos
- Proporcionar soporte para más requerimientos que los detectados inicialmente

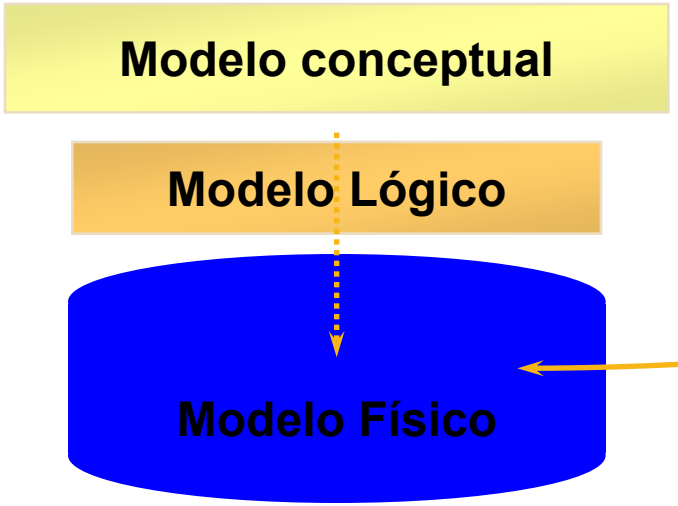


Modelado multidimensional

El proceso de modelado de un DW

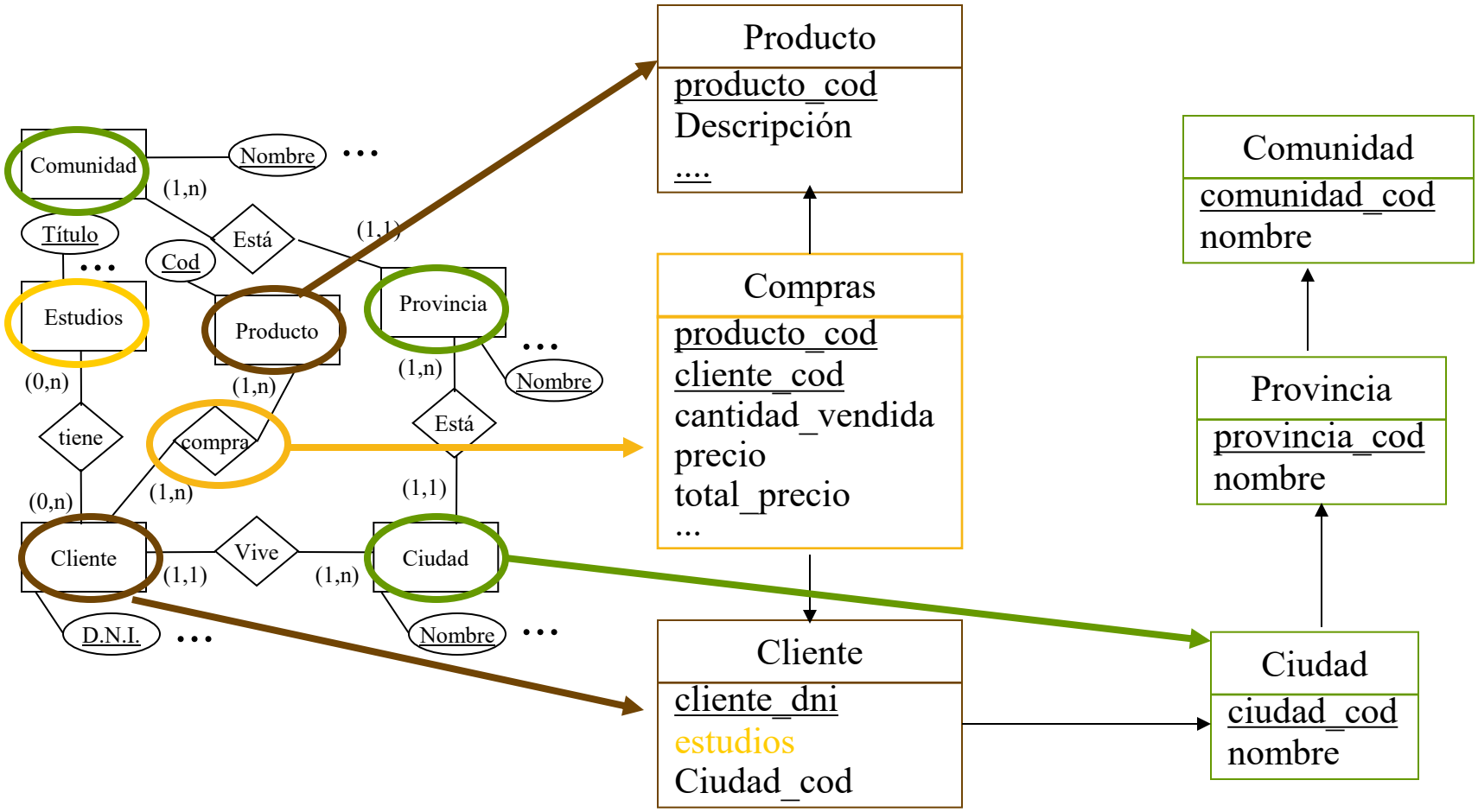
- Crear un modelo de negocio
  - Opcional
- Crear un modelo dimensional
- Crear un modelo físico





Modelado multidimensional

El proceso de modelado de un DW. Modelo EER → Esq. Estrella



Bibliografía

- Giovinnazo (2000). Object-Oriented Data Warehouse Design: Building a star schema
- Inmon (2002). Building the Data Warehouse (3ª ed.)
- Kimball (2002). The Data Warehouse Toolkit (3ª ed.)
- Thomsen (2000). OLAP solutions: Building Multidimensional Information Systems

Diseño lógico de almacenes de datos

Tema 3

Profesores:

Juan C. Trujillo
Alejandro Reina Reina
LUCENTIA Research Group

Universitat d'Alacant
Universidad de Alicante

Departamento de Lenguajes y Sistemas Informáticos

Tema 4: Preparación de datos en el contexto de la IA

Adquisición y preparación de datos

mpilar.escobar@ua.es

Universitat d'Alacant
Universidad de Alicante

Tema 4: Preparación de datos en el contexto de la IA

Contextualizar

- Ciclo vida proyecto IA
- Fases de la preparación de datos
- Adquisición de datos
- Introducción al diseño de procesos ETL con Pentaho

¿Qué es la preparación de los datos?

¿Por qué es importante?

¿Qué hace que los datos estén preparados para la IA?

Preparación de datos

Proceso: conjunto de pasos necesarios para la transformación de los datos en un formato adecuado, limpios y de calidad para que puedan ser utilizados por los modelos de IA.

Tareas clave: recolectar, limpiar, estructurar y enriquecer los datos.

Objetivo es garantizar la calidad, la coherencia y la relevancia de los datos para maximizar el rendimiento y la fiabilidad de los modelos de IA.

El 30% de los proyectos de IA generativa serán abandonados tras la prueba de concepto debido a la baja calidad de los datos, a controles de riesgo insuficientes, al aumento de los costes o a un valor empresarial poco claro (Gartner).

Preparación de datos

¿Están preparados para los modelos de IA?

Datos preparados para la IA

Alineación	Calificación	Gobernanza
• Cuantificación • Semántica • Calidad • Confianza y equidad • Diversidad • Linaje	• Evaluación de la coherencia • Validación y verificación • Acuerdos de nivel de servicio (SLA) operativos • Control de versiones • Pruebas de regresión continuas • Indicadores de observabilidad	• Administración de datos • Inferencia y derivación • Regulación y cumplimiento • Apoyo a las normas de IA • Intercambio de datos

Fuente: Gartner

Ciclo de vida del proyecto IA

- Objetivos del modelo
- Casos de uso
- Analizar y evaluar necesidades de datos

- Recolectar
- Limpiar
- Estructurar
- Enriquecer datos

- Selección del modelo
- Entrenamiento
- Pruebas, ejecución y optimización del modelo

- Métricas clave
- Validar fiabilidad
- Alineación con los objetivos

Ciclo de vida del proyecto IA

Etapa 1: Definición del proyecto centrado en los datos

Contextualización
de los datos

Ciclo de vida del proyecto IA

Etapa 1: Definición del proyecto centrado en los datos

Determinar el alcance del problema a resolver.

Lienzo del problema (4W)

- ***Who***, quiénes son las partes interesadas
- ***What***, cuál es el problema o necesidad y cuál es su naturaleza
- ***Where***, la situación, momento y contexto dónde se produce
- ***Why***, cuál es el beneficio y su impacto, tanto en las partes interesadas como en la sociedad

Ciclo de vida del proyecto IA

Etapa 1: Definición del proyecto centrado en los datos

Determinar el alcance del problema a resolver

Definir los objetivos y los casos de uso.

Analizar las necesidades a nivel de datos:

- Qué datos son necesarios: tipo, formato …
- Qué fuentes de datos se van a utilizar

Definir métricas clave y evaluar en función de los objetivos y casos de uso:

- Expectativas de los casos de uso: cuantificación, semántica, calidad, confianza y diversidad
- Requisitos de confianza: validación, rendimiento, coste y otros requisitos no funcionales
- Requisitos de gobernanza de datos asegura que los datos y los modelos se gestionen de forma legal, ética y eficiente

Ciclo de vida del proyecto IA

Etapa 2: Adquisición y preparación de datos

Pasos para la preparación de los datos:

1. Adquisición de datos
2. Limpieza de datos: unificar, reducir …
3. Transformación de datos
4. Normalización de los datos
5. Etiquetado de datos
6. Enriquecimiento de datos
7. Validación de datos
8. Equilibrar el conjunto
9. Particionamiento de datos

Ciclo de vida del proyecto IA

Etapa 3: Modelización. Proceso iterativo

- Selección/desarrollo del modelo
- Entrenamiento usando los datos preparados en la fase anterior
  - Pruebas y ejecución para tratar de predecir comportamientos o descubrir ideas en los datos
- Afinado del modelo para resolver el problema planteado
  - Proceso iterativo →retroalimentación con los datos
  - Alineación con los objetivos del proyecto
- Evaluación de otros modelos
  - Precisión, robustez, eficacia, etc.

Ciclo de vida del proyecto IA

Etapa 4: Evaluación

Evaluar en función de las métricas definidas, por ejemplo:

- Precisión
- Recuperación (sensibilidad)
- Fiabilidad del modelo

Alineación con los objetivos definidos previamente y su impacto.

Responsabilidad ética y legal.

Ciclo de vida del proyecto IA

Etapa 5: Despliegue en producción

No todos los proyectos de IA pueden o deben ser puestos en producción.

El coste de poner un modelo en producción puede ser elevado incluso superar el valor que aportaría.

Hacer previsión antes de construir el modelo, aunque esto no siempre es posible.

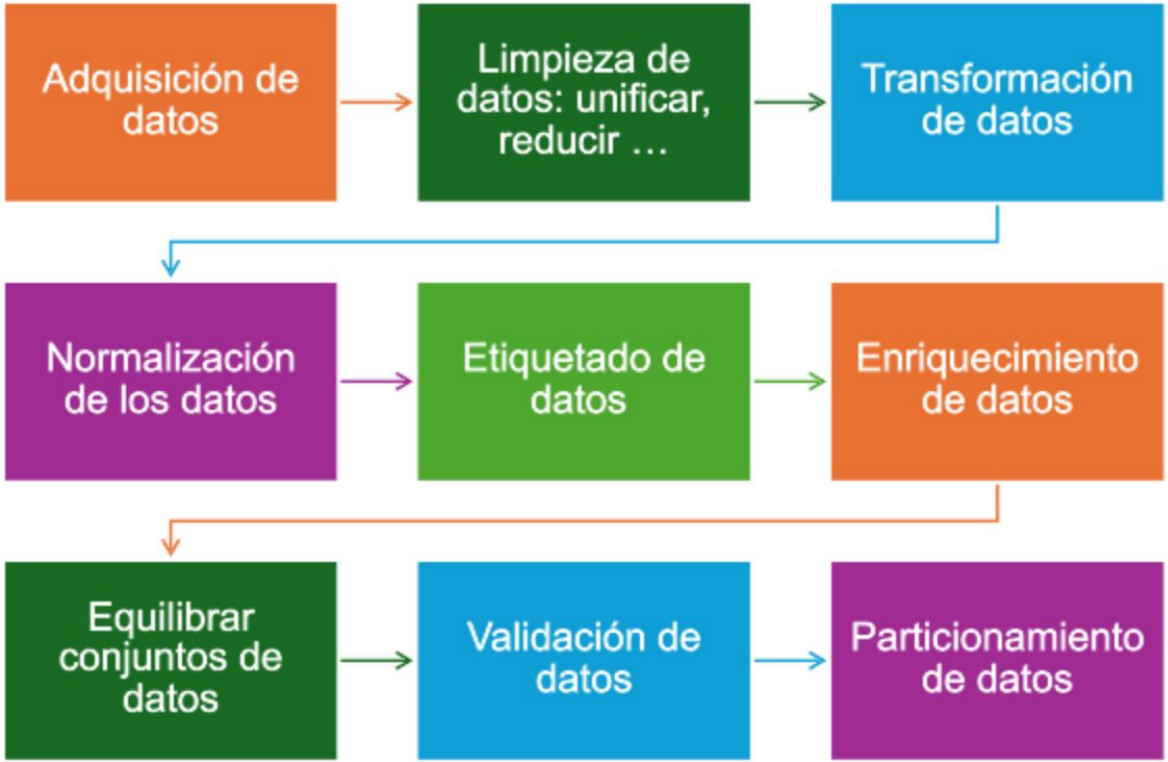
Monitorizar modelo comprobando que siga siendo preciso y eficaz y se adapta a los nuevos datos y cambio en las condiciones.

Ciclo de vida del proyecto IA

Retos más comunes

- Escasez o desequilibrio de datos
- Sesgo en datos y modelos
- Cumplimiento y normativa
- Sobreajuste y pérdida de generalizabilidad
- Latencia en aplicaciones de tiempo real
- Complejidad de la integración
- Gestión de costes

Preparación de datos

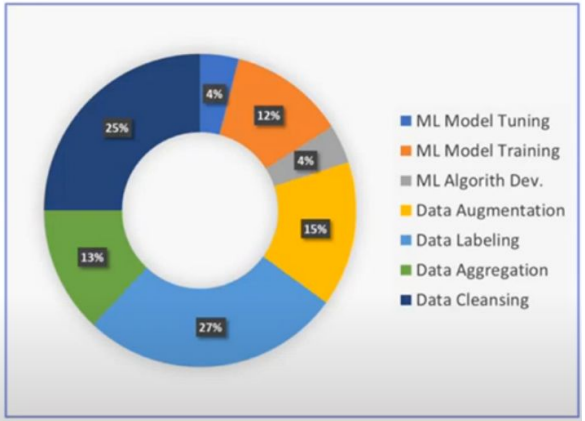


Preparación de datos

3. PREPARACIÓN DE DATOS

SIGMA • COGNITION

- 80% del tiempo en proyectos de IA
- Garantizar que los datos que alimenten a los modelos de AI cumplan los siguientes requisitos:
  - ✓ Limpios
  - ✓ Precisos
  - ✓ Completos
  - ✓ Relevantes



Fuente: https://www.cognilytica.com

Adquisición de datos

Objetivo: Obtener datos relevantes y confiables para el entrenamiento de los modelos de IA.

Aspectos relevantes en la recolección de datos:

- Identificar las necesidades del proyecto
- Seleccionar fuentes de datos adecuadas y confiables
- Seleccionar datos relevantes, representativos y diversos
- Verificar el cumplimiento legal y ético FAIR-R + CARE
- Automatizar la recolección si es necesario
- Evaluar la calidad de los datos recopilados

Adquisición de datos

Seleccionar fuentes de datos, variedad y veracidad

Fuentes externas			**Fuentes internas**
***Datasets* públicos**	***Datasets* privados**	**Datos generados**	
Datos abiertos, datos abiertos enlazados y API públicas Licencia Creative Commons Primeras aproximaciones del modelo Adaptaciones más ligeras al dominio Se puede aplicar a datos experimentales	Plataformas de datos de terceros. Normalmente bajo licencia o acuerdo con el responsable Administración propia de los datos Adaptación más cercana al dominio y situación real donde operará el modelo.	Capturas de sensores, datos de IoT o simulaciones. Comprobar la credibilidad y la puntualidad de estas fuentes para garantizar que los datos sean pertinentes y precisos	BD corporativas, registros de usuarios o historiales Administración propia de los datos, coste añadido Cubre casuística Datos personalizados según las necesidades

Adquisición de datos

Clasificación de los datos según formato y esquema

	Datos Estructurados	Datos No Estructurados	Datos Semiestructurados
Estructura	Modelo de datos predefinido y estricto	Sin modelo de datos predefinido	Estructura más flexible y jerárquica Uso de metadatos
Almacenamiento	Formato tabular Esquema rígido bases de datos relacionales y almacenes de datos*	Formato nativo, bases de datos NoSql, data lakes	Bases de datos NoSQL
Análisis	Fácil y rápido	Más complejo y requieren habilidades y herramientas especializadas	Moderadamente complejo
Ejemplos	Tablas, bases de datos Sql, hojas de cálculo, csv*	Textos, imágenes, videos, multimedia, IoT, social media	JSON, XML
Casos de uso	Machine learning (ML)	PLN Modelos IA generativa RAG (Generación Aumentada por Recuperación) Análisis predictivos	

Adquisición de datos

Seleccionar fuentes de datos, datos abiertos y **FAIR-R**

- Ámbito de datos abiertos y ciencia abierta
- FAIR Enfocado en las características del dato

Aspectos de FAIR-R

- Etiquetado, anotado y enriquecimiento
- Transparencia sobre el origen, linaje y tratamiento de los datos
- Estándares, metadatos, esquemas y formatos óptimos para su uso por parte de la IA
- Cobertura y calidad suficientes para evitar sesgos o falta de representatividad

Adquisición de datos

Seleccionar fuentes de datos, datos abiertos y CARE

- Ámbito de datos abiertos y ciencia abierta
- Complementan FAIR considerando tanto a las personas como al propósito

CARE Principles for Indigenous Data Governance

PRINCIPIOS CREA

DATOS INDIGENAS

Control (autoridad para controlar)
Responsabilidad
Ética
Aprovechamiento para el beneficio colectivo

Adquisición de datos

Requisitos para la adquisición de los datos

Datasets existentes	**Anotación de datos**
Dominio de la aplicación Representa la variabilidad Clases están balanceadas	Calidad Regulación ética y legal (GDPR, LOPD, CARE) Seguridad Escalabilidad

Adquisición de datos

Ejemplos de fuentes de datos

- Data science community https://www.kaggle.com/
- World Bank o Banco Mundial World Bank Open
- datos.gob.es
- European data
- FAO catalog datasets
- Geonames.org
- Copernicus EU
- TData CV datos sobre el turismo
- INE
- Open food facts
- AEMET OpenData
- OpenWeatherMap

Adquisición de datos

Casos de uso

- Google Dataset Search
- DataMed
- Copernicus
- Administraciones públicas:
  - datos.gob.es, guía práctica para la publicación de datos abierto usando APIs
  - EDP dataset search interface Portal europeo de datos
  - https://data.europa.eu/es/publications/use-cases
  - Open Data Barcelona → Ejemplo RDF Data Cube
- Patrimonio cultural
  - Bibliotecas digitales CBVMC
  - British Library datasets
  - Gallica, la biblioteca digital francesa
  - Library of Congress

Procesos ETL

Procesos ETL

En el contexto de la IA los procesos ETL son una opción para la adquisición y preparación de los datos para los modelos.

Herramienta de integración de datos

Facilita:

- Adquisición y gestión de datos
- Gestión y administración de servicios
- Extracción de datos
- Transformación de datos
- Carga de datos en el sistema destino

Procesos ETL

Extract

- Extraer datos desde las distintas fuentes de origen - fusión de datos
  - Identificar fuentes
  - Identificar conjunto de datos
  - Conectores para la adquisición
  - Fuentes como Hadoop, datos abiertos, Sql, Nosql (MongoDB, Cassandra), repositorios RDF, Wikidata, …
- Analizar los datos obtenidos
  - Tipos de datos en qué formato origen
  - Codificación origen
- Interpretar y comprobar que cumplen con la estructura esperada
- Convertir al formato más adecuado para nuestro objetivo

Procesos ETL

Transform

En esta fase se aplican las reglas de negocio que guiarán los datos obtenidos aplicando transformaciones para el formato final, datos en contenido y forma

- Selección de datos (columnas, filas…) - Solo lo necesario el resto es ruido
- División de columnas según necesidades
- Traducción de valores
- Codificación
- Combinación de fuentes
- Combinación de datos para obtener nuevos
- Tratamiento de textos (operaciones sobre ellos expresiones regulares)

Procesos ETL

Load

Cargar los datos obtenidos y transformados en el sistema destino

- Selección sistema destino: fichero, BD relacional, índice, NoSql
- Nuevo o sobrescribir (ejemplo BD, Excel)
- Historial de cambios

Tipos de proceso de carga:

- Acumulación simple
- Rolling - mayor granularidad

Flujos de trabajo de procesos ETL

Existen diferentes herramientas que permiten trabajar con grandes volúmenes de datos y facilitan su análisis.

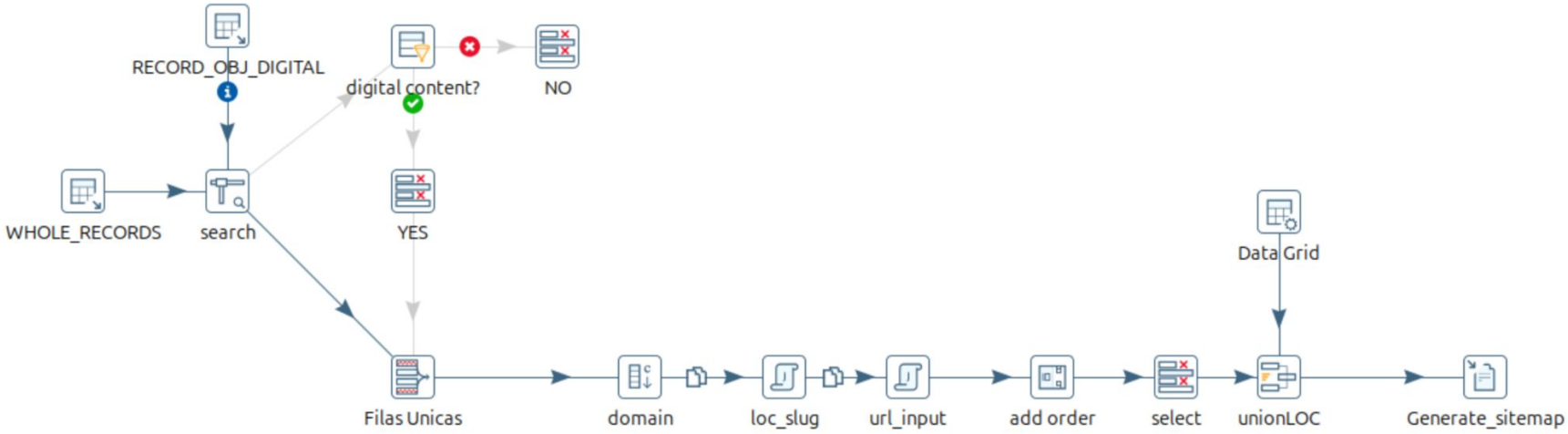
Para ello se diseñan y desarrollan flujos de trabajo que se pueden crear utilizando pasos o entradas unidas por saltos que pasan datos de un elemento al siguiente.

Los **flujos de trabajo** utilizan dos tipos de archivos:

1. Transformaciones que realizan tareas ETL.
2. Trabajos que organizan actividades de ETL, como la definición del flujo, dependencias y ejecución.

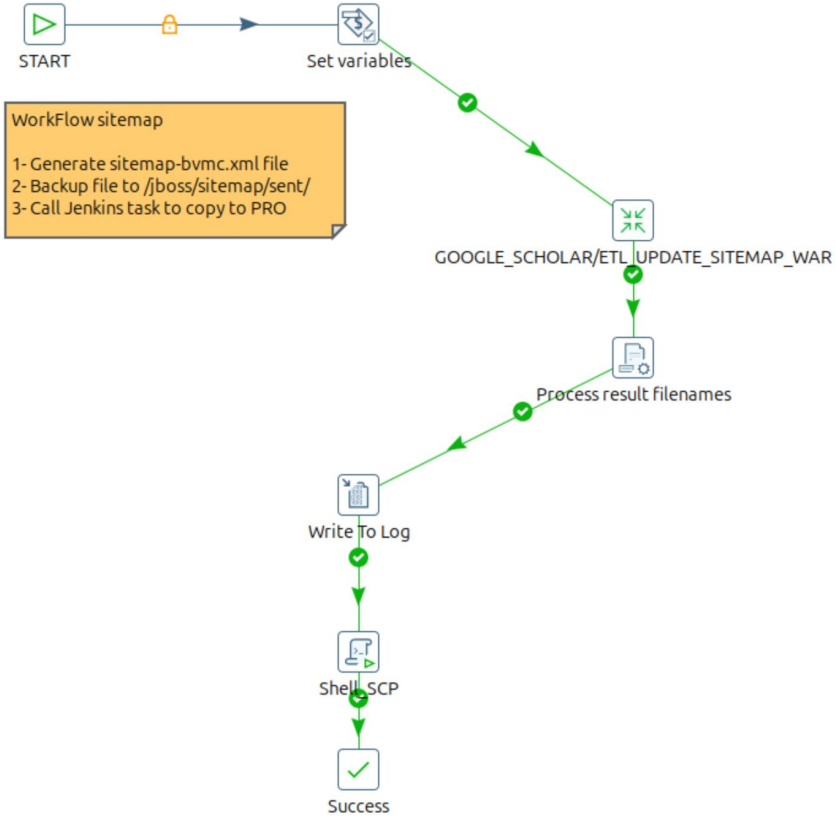
Flujos de trabajo de procesos ETL

Transformation



Flujos de trabajo de procesos ETL

Job



Pentaho, herramientas para diseño de ETLs

Pentaho suite Business Intelligence para los procesos de gestión de datos.

Open source multiplataforma creada en 2004, plataforma Java.

Disponible en dos ediciones: **Community Edition** (CE) y **Enterprise Edition** (EE).

Múltiples plugins creado por la comunidad de código abierto.

Pentaho Community wiki
https://pentaho-public.atlassian.net/wiki/spaces/COM/overview?mode=global

Pentaho, herramientas para diseño de ETLs

Pentaho Data Integration (Kettle)

Entorno gráfico para diseño flujos de datos.

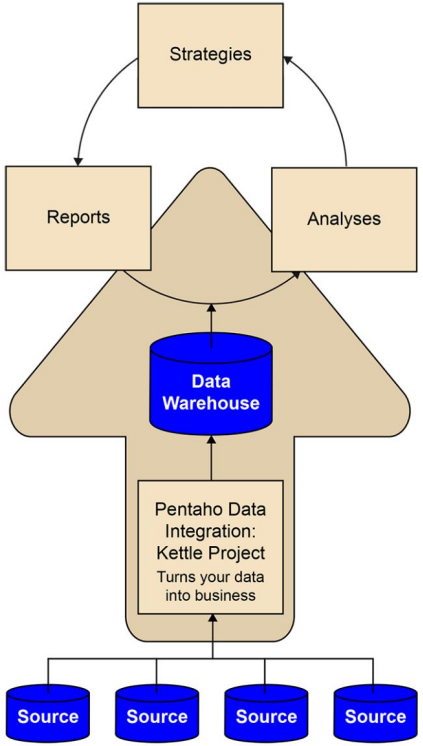
Servicio de integración de datos ETL (extract, transform, load).

Facilita el acceso a una amplia gama de fuentes de datos, incluidas bases de datos de código abierto y propietarias.

Arquitectura extensible mediante plugins y perspectivas

Algunos usos de Pentaho Data Integration

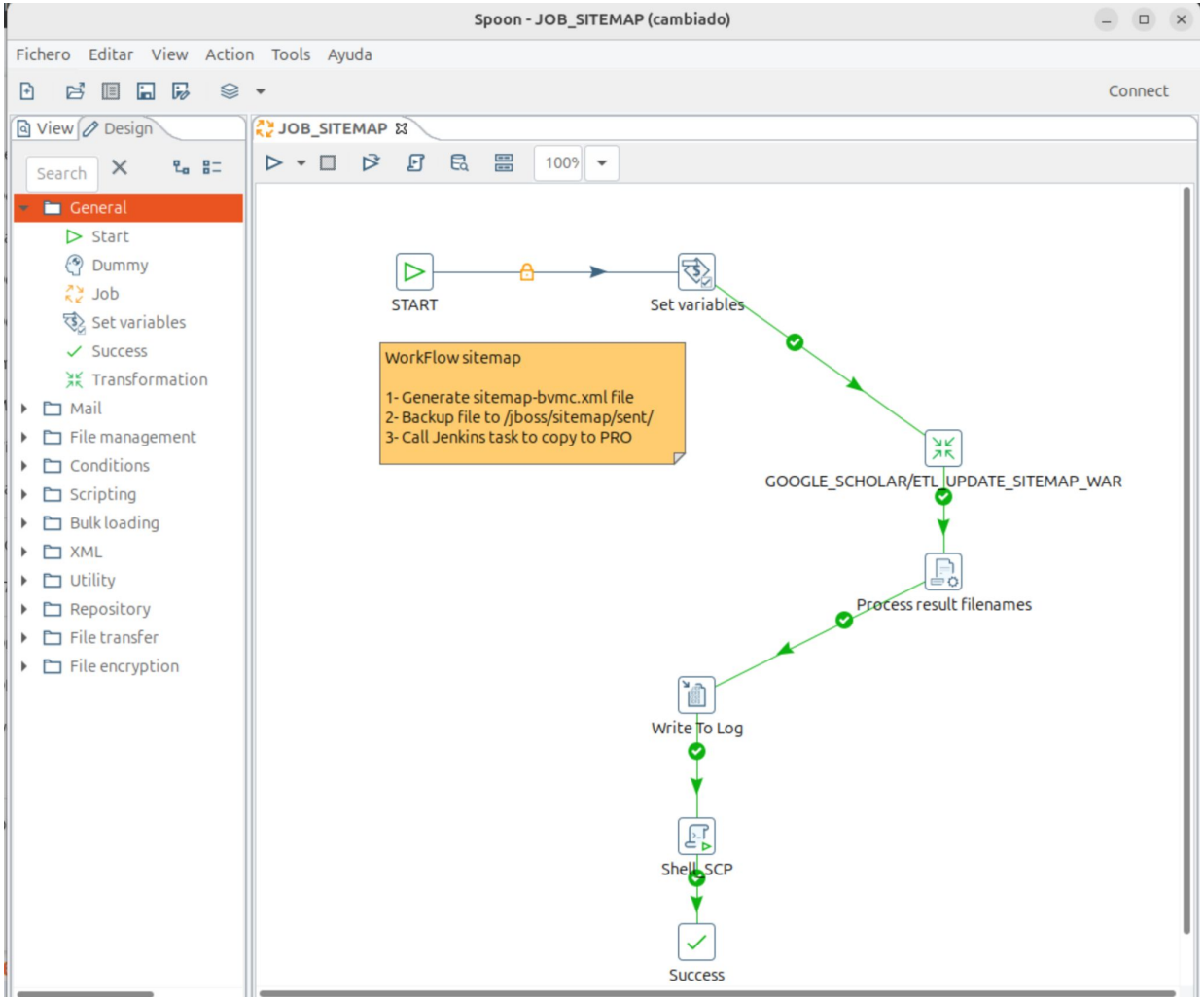
- Migración de datos entre diferentes sistemas
- Carga de grandes volúmenes de datos
- Limpieza de datos mediante transformaciones
- Integración de datos mediante herramientas ETL en tiempo real



Fuente: Pentaho.

Pentaho

Entorno visual



Pentaho, herramientas para diseño de ETLs

Componentes clave de Pentaho Data Integration

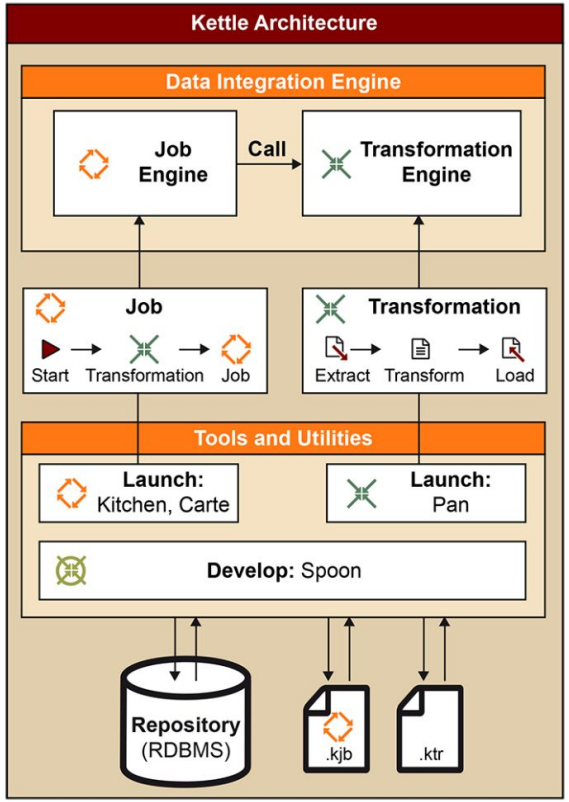
Spoon: entorno gráfico para creación de flujos de trabajo, transformaciones y trabajos.

- Realizar data warehouse
- Operaciones, como crear conexiones de datos, transformaciones o inserción de fórmulas, entre otras.
- Múltiples funcionalidades para realizar procesos de ejecución, transformación y carga de datos.

Pan: creación líneas de comandos para ejecutar transformaciones

Kitchen: ejecución de trabajos por línea de comandos

Carte: servidor web para ejecutar transformaciones y trabajos

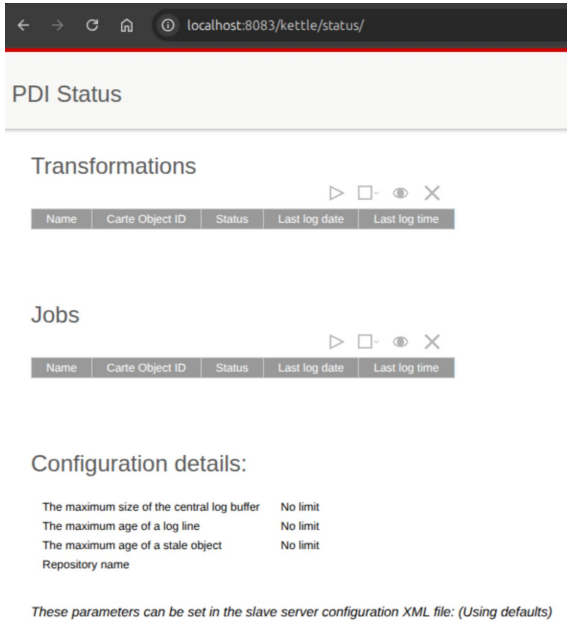


Fuente: Pentaho.

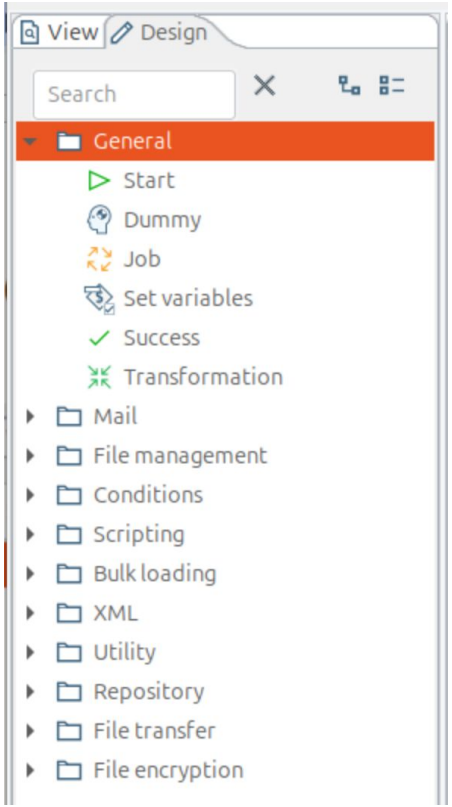
```
carte.sh*
encr.sh*
import.sh*
kitchen.sh*
pan.sh*
purge-utility.sh*
runSamples.sh*
set-pentaho-env.sh*
SpoonDebug.sh*
spoon.sh*
yarn.sh*
```

```
#!/bin/bash

bash ~/Aplicaciones/Pentaho/data-integration/kitchen.sh -file
/home/GS/JOB_SITEMAP.kjb -level:Basic

bash ~/Aplicaciones/Pentaho/data-integration/pan.sh -file:/home/ETL_GEN_ARK.ktr
-level:Basic
```

bash ./**carte**.sh 127.0.0.1 8083



Pentaho Data Integration



Pasos para el diseño de *jobs*

General, pasos de inicio de trabajo, ejecución transformaciones o trabajos entre otras operaciones

Mail, envío de correos, recuperación y validación de cuentas

File management, gestión de operaciones sobre ficheros y carpetas, HTTP, etc

Conditions, comprobaciones sobre conexiones, BD, ficheros, etc

Scripting, scripts en JavaScript, Shell y SQL

Bulk loading, cargas a MySQL y MSSQL

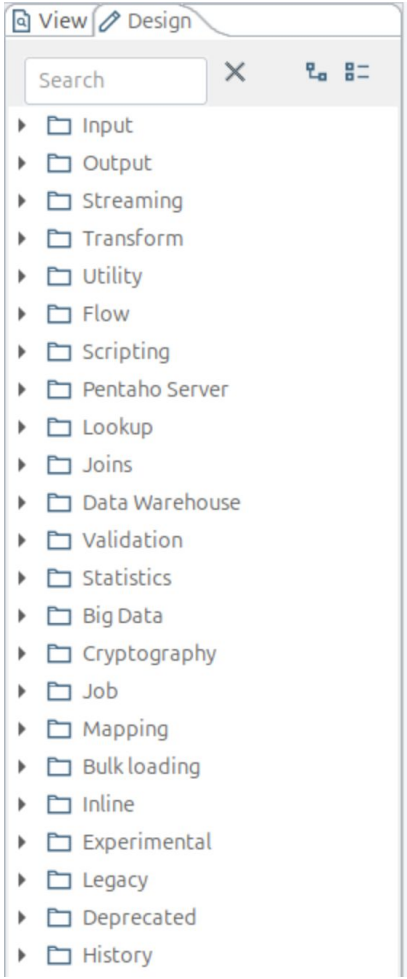
XML, validaciones XML, DTD, XSD y XSL

Utility, pasos para ejecución de transformaciones como ping, enviar información a syslogs, Nagios, etc

Repository, operaciones con el repositorio de transformaciones y trabajos.

File transfer, gestión ficheros FTP, FTPS y SFTP

File encryption : sistema PGP (Pretty Good Privacy) para envío y recepción de ficheros



Pentaho Data Integration

Pasos para el diseño de *transformations*

Input, adquisición de datos desde diversas fuentes, BD, CSV, Excel, JSON, repositorios, etc

Output, carga datos en diferentes sistemas

Streaming, obtener datos de Kafka, JMS (Java Message Service) o MQTT (Standard for IoT Messaging)

Transform, ejecución de diversas operaciones con datos, como filtrar, ordenar, partir, añadir nuevos campos, mapear, etc

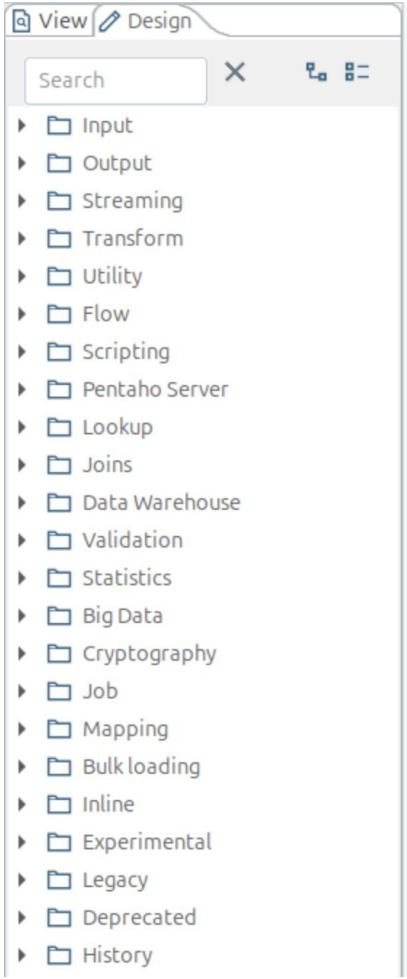
Utility, ejecución ssh commands, envío de mensajes syslog, etc

Flow, control de flujo de datos en las transformaciones

Scripting, SQLScript, JavaScript, expresiones regulares, fórmulas y expresiones java

Pentaho server, conexión al servidor de Pentaho para operaciones como la gestión de variables

Lookup, pasos para añadir información al flujo de datos mediante búsqueda en fuentes como BD, web services, HTTP, etc



Pentaho Data Integration

Pasos para el diseño de *Transformations*

Joins, unión de filas en función de diferentes criterios

Data Warehouse, para trabajar con dimensiones

Validation, validaciones como tarjetas de crédito, datos, email o XSD

Statistics, operaciones estadísticas sobre un flujo de datos como group by

Big Data, cargar y extraer datos de Avro y MongoDB

Cryptography, uso de PGP (Pretty Good Privacy)

Job, realizar operaciones propias de un Job

Mapping, mapeo entre campos de entrada y salida

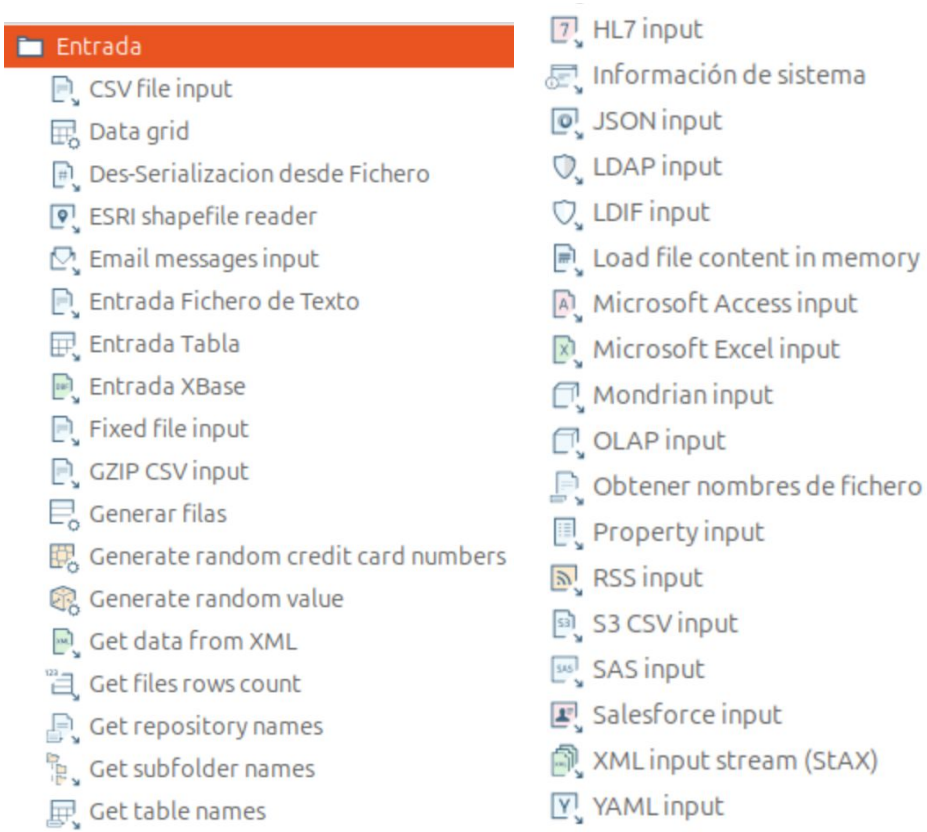
Bulk loading, cargas desde Vectores, MySQL, Oracle, PostgreSQL entre otras

Inline, embebido para operaciones con sockets

Experimental, pasos en fase de validación

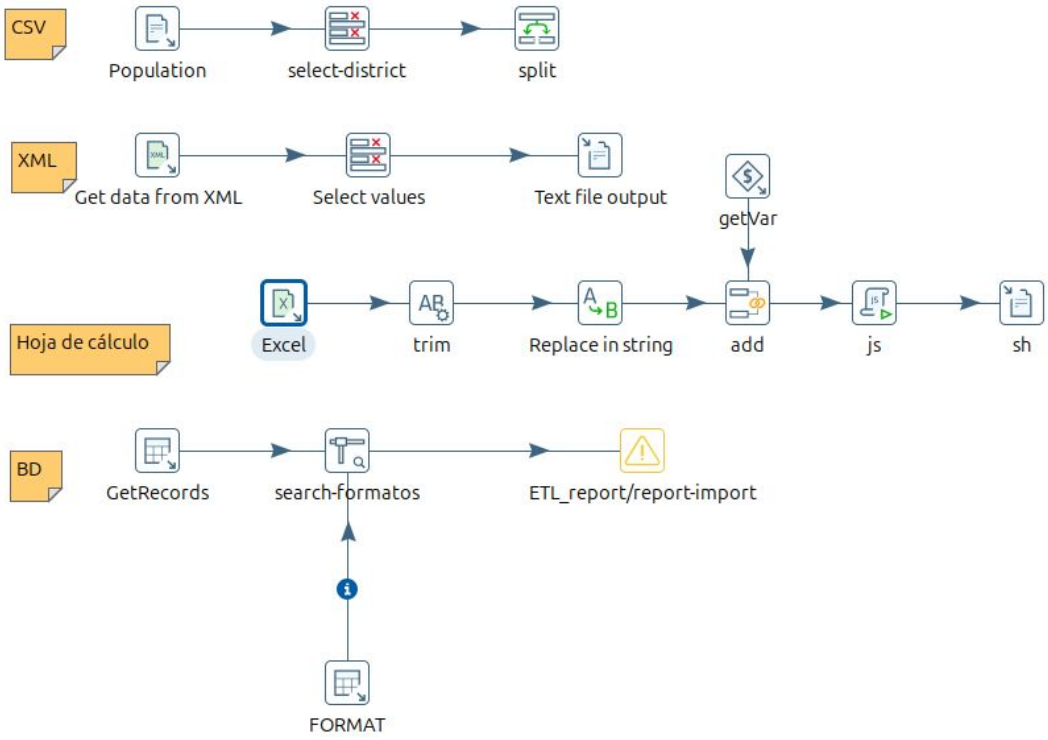
Pentaho Data Integration

Extract, adquisición de datos y distintos conectores para el acceso a las fuentes.



Pentaho Data Integration

Ejemplo 1: conexión a distintas fuentes de datos



Tema 4: Caso práctico

Con el objetivo de entender el diseño de procesos ETL, se propone un caso práctico real, pero simplificado. Donde la explotación y reutilización de los datos proporcionados por plataformas Open Data hace posible realizar previsiones, adelantándose a posibles problemas y minimizando así su impacto. En muchos casos es posible llegar a evitarlos.

Se utilizará datos abiertos publicados por la plataforma Open Data BCN en el contexto de Open Government Data. Este escenario se centra en el estado de los puntos críticos de limpieza en esta ciudad. En primer lugar, se realizó un análisis de los datos disponibles en esta plataforma seleccionando archivos sobre entorno urbano y estado de las áreas críticas de limpieza en la ciudad de Barcelona así como información sobre límites administrativos.

A continuación se listan los conjuntos de datos seleccionados para esta experimentación, junto con el formato en el que estaban disponibles originalmente en la fecha en la que se realizó la prueba real.

- renta-poblacion.csv datos sobre la distribución de los ingresos y la población en la ciudad
- edm.csv datos sobre la población en la ciudad
- puntscriticsbarcelona.csv información sobre el entorno urbano y estado de las áreas críticas de limpieza en la ciudad de Barcelona así como información sobre límites administrativos

*Algunos datos estaban en formato texto y se transformaron a CSV, se ha simplificado para la realización de este ejemplo

Tema 4: Caso práctico

La estrategia en el diseño del proceso ETL será la siguiente:

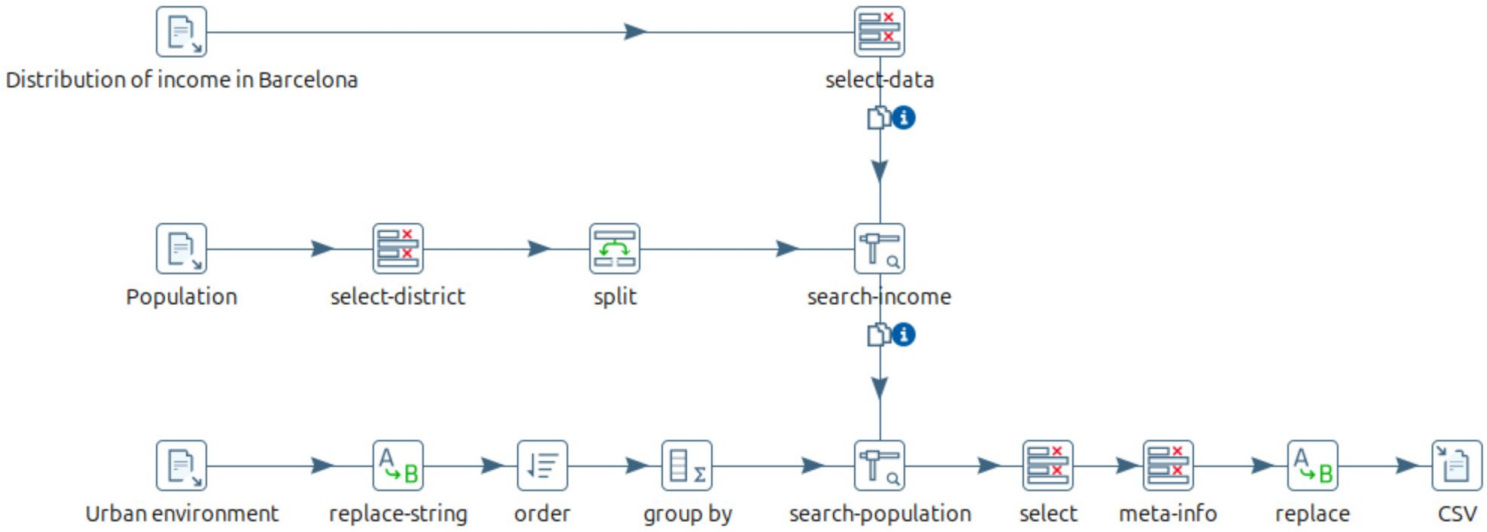
1. Cargar los datos a partir de los ficheros anteriores seleccionando los datos a utilizar en cada caso.
2. Se identifican los conjuntos de datos que comparten puntos comunes (distrito, ubicación geográfica, etc.) y por tanto, permiten un análisis más detallado.
3. Complementar las dimensiones sobre información urbana a partir de la información del fichero edm.csv y el de renta-poblacion.csv.
4. Crear un trabajo para lanzar todas las transformaciones de una manera única.

Notación utilizada:

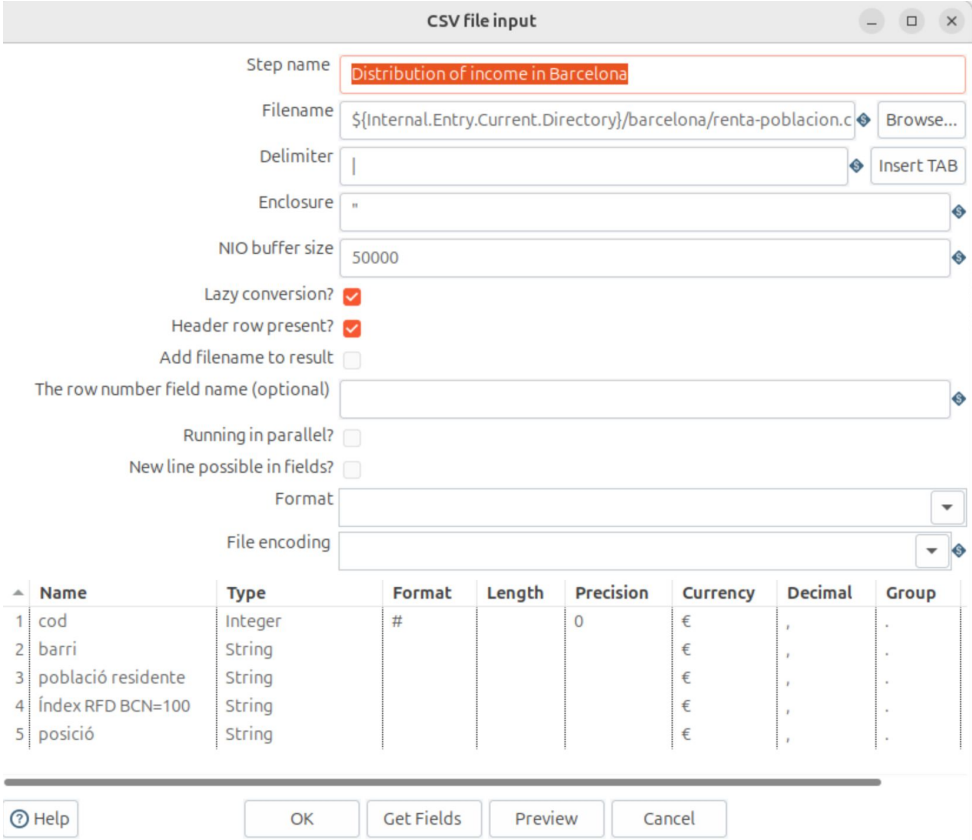
Para las transformaciones:ETL_INI_{Nombre de la dimensión, tabla de hecho que cargar o funcionalidad}.

Para los trabajos: JOB_INI_{Nombre del proceso}.

Tema 4: Caso práctico



Tema 4: Caso práctico



Tema 4: Caso práctico

CSV file input

Step name: Population

Filename: ${Internal.Entry.Current.Directory}/barcelona/edm.csv [Browse...]

Delimiter: , [Insert TAB]

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☑

Header row present? ☑

Add filename to result ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

Format:

File encoding:

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Dte.	String				€	,	.	both
2	Barri de baixa	String				€	,	.	both
3	Districte d'alta: Mitjana edat homes	Number	#.#		0	€	.	,	none
4	Districte d'alta: Mitjana edat dones	Number	#.#		0	€	.	,	none
5	Districte de baixa: Mitjana edat homes	Number	#.#		0	€	.	,	none
6	Districte de baixa: Mitjana edat dones	Number	#.#		0	€	.	,	none

Help | OK | Get Fields | Preview | Cancel

Tema 4: Caso práctico

CSV file input

Step name: Urban environment

Filename: ${Internal.Entry.Current.Directory}/barcelona/puntscriticsbarcelona.csv [Browse...]

Delimiter: , [Insert TAB]

Enclosure: "

NIO buffer size: 50000

Lazy conversion? ☑

Header row present? ☐

Add filename to result ☐

The row number field name (optional):

Running in parallel? ☐

New line possible in fields? ☐

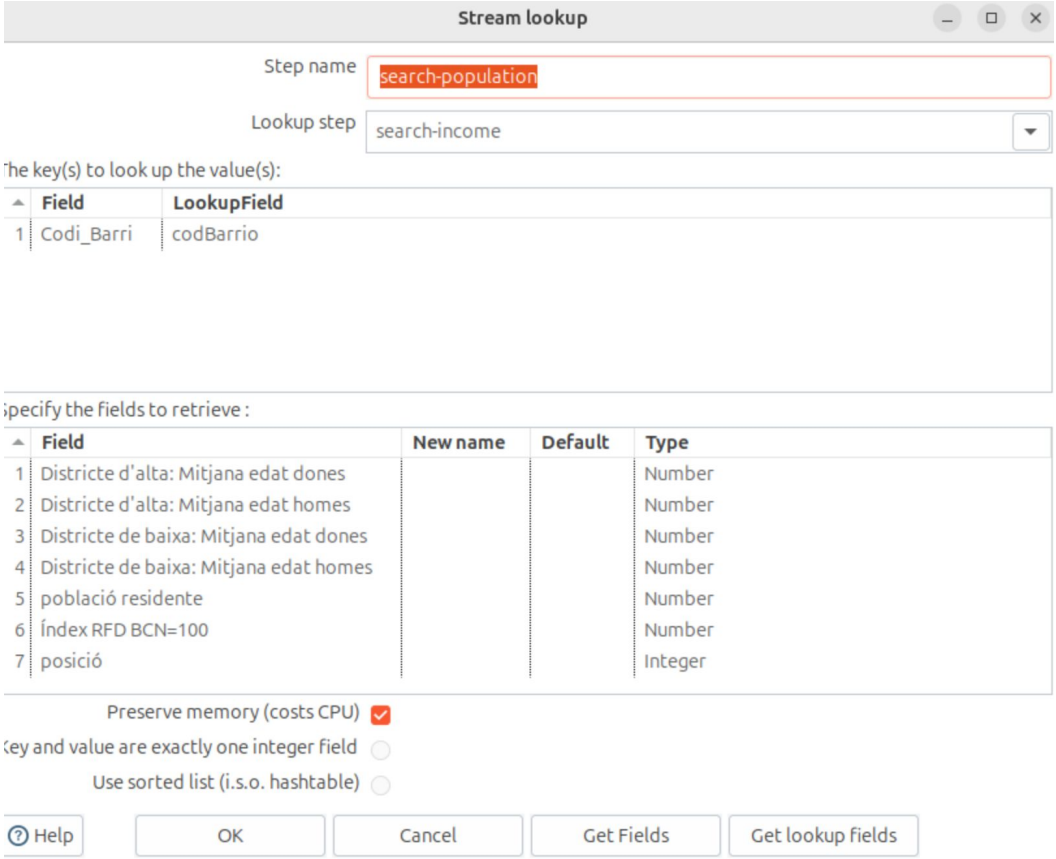
Format:

File encoding:

	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	Número_punt_crític	Integer	#		0	€	,	.	none
2	Codi_Districte	Integer	#		0	€	,	.	none
3	Nom_Districte	String				€	,	.	none
4	Codi_Barri	Integer	#		0	€	,	.	none
5	Nom_Barri	String				€	,	.	none
6	Nom_Carrer	String				€	,	.	none
7	Número	String	#		0	€	,	.	none
8	Latitud	String			0	€	.	,	none
9	Longitud	String			0	€	.	,	none

Help | OK | Get Fields | Preview | Cancel

Tema 4: Caso práctico



Tema 4: Caso práctico

Extracto de los datos que contiene el fichero con información sobre los puntos críticos de limpieza en la ciudad de Barcelona. Datos extraídos de Open Data BCN.

```
Martina Castells ;1;41.379772;2.166827;Mala praxis comercial;20;15.8;18;7.4;18;10.8;11;4.3;18;4.2
Malnom ;1;41.380627;2.167949;Problemàtica social;19;4;18;4.9;18;3.3;16;4.6;19;3.7
Picalquers ;1;41.380665;2.168612;Problemàtica social;19;12.1;18;8;18;12.7;16;13.1;19;12.3
Rubió i Luch;1;41.38073;2.170163;Problemàtica social;21;6.9;19;4.3;18;4.7;16;3.2;19;2.9
Egipcíaques;3;41.38089;2.168989;Mal ús de contenidors i/o papereres;19;7.6;19;9.3;18;10.4;16;7.3;20;6.9
Gardunya ;1;41.381252;2.17052;Força afluència de gent;19;7.1;19;5.8;17;5.3;16;4.1;19;5.4
Sant Agustí;0;41.38073;2.171528;Mala praxis comercial;19;9.2;19;14.6;18;6.2;16;6.7;19;11.9
Robador;53;41.378761;2.171453;Efecte crida;23;3.4;19;3.6;18;5.8;16;6.4;20;8.3
Salvador Seguí;1;41.37878;2.170497;Mal ús de contenidors i/o papereres;23;5.3;19;5.5;18;3.6;16;4.1;20;7.6
Aurora ;11;41.378757;2.168208;Efecte crida;23;7.3;18;11.1;18;15.1;16;18.5;19;16.1
Aurora ;25;41.377934;2.167251;Efecte crida;23;10.8;18;7.5;18;16.1;16;15.7;19;18.9
Sant Rafael ;42;41.378326;2.1683;Mal ús de contenidors i/o papereres;23;13.3;17;14.8;18;17.1;16;17.1;19;17.8
```

```
01|el Raval|47.274|74,6|48|
38|la Teixonera|11.281|75,8|47
02|el BarriGòtic|15.729|110,5|16|
39|Sant Genís dels Agudells|6.828|86,3|33
03|la Barceloneta|15.068|84,8|35|
40|Montbau|5.101|82,2|41
04|Sant Pere, Santa Caterina i la Ribera|22.380|97,8|24|
41|la Vall d'Hebron|5.687|87,5|32
```

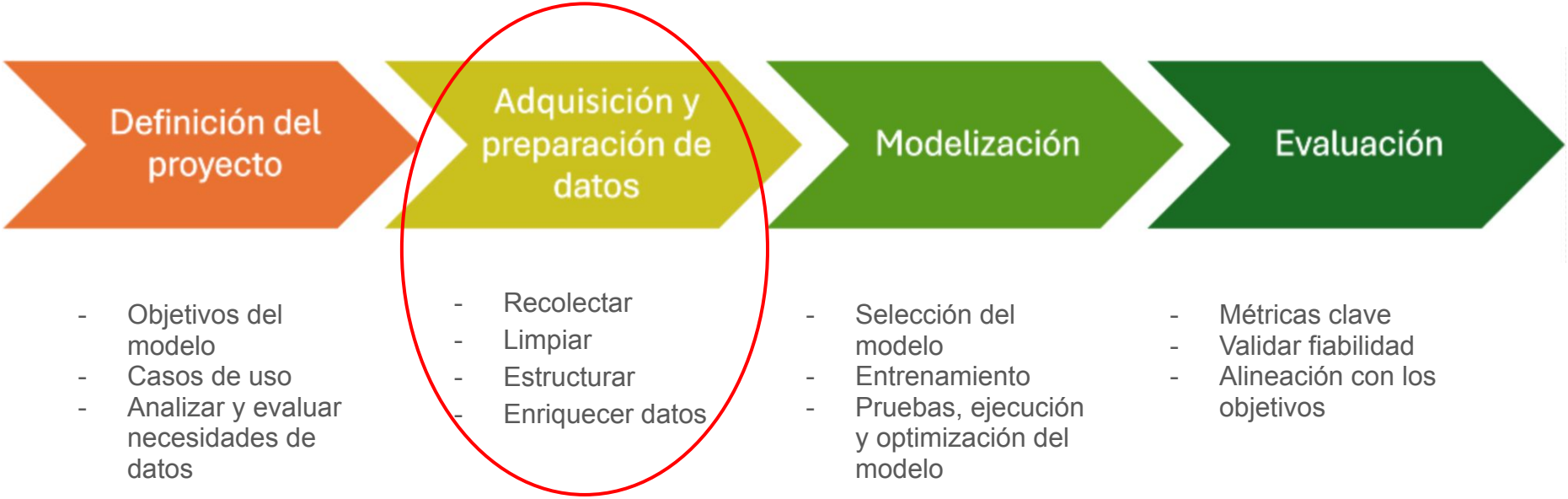
Procesos ETL

Tema 5: Introducción a la integración de datos mediante ETLs. Parte I

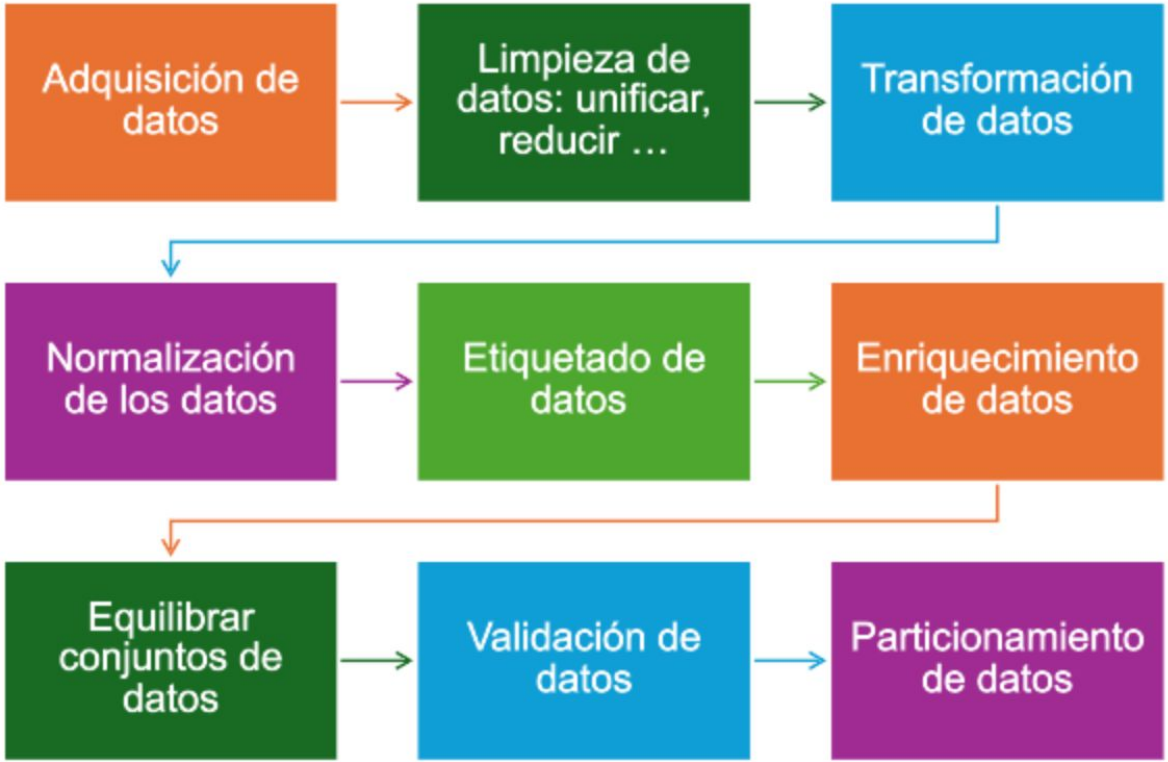
Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Ciclo de vida del proyecto IA



Preparación de datos



Tema 5: Introducción a la integración de datos mediante ETLs

Parte I - Integración de datos

- Qué es la integración de datos
- Técnicas de integración
- Integración con ETL en Pentaho, componentes, flujos de datos, tipos de datos, conversiones…

Tema 5: Introducción a la integración de datos mediante ETLs

Parte II - Extracción de datos a través de procesos ETL

- Database-Based Extraction
- File-Based Extraction
  - Working with CSV/Excel
  - Working with XML files
- Web-Based Extraction
  - Web services
  - HTTP Client
  - Text-Based Web Extraction

Tras definir el proyecto y seleccionar las fuentes de datos, el siguiente paso es la carga y preparación de esos datos para que puedan ser usados.

¿Cómo?

A menudo nos enfrentamos a desafíos como:

- **Fuentes heterogéneas,** formatos diferentes como SQL, archivos CSV, XML, RDF, JSON acceso a APIs, etc.
- **Calidad de los datos**, los datos de las diferentes fuentes pueden tener errores, estar incompletos o ser inconsistentes.
- **Gestión de gran volumen de datos** que se generan a gran velocidad.
- **Seguridad y privacidad**, cómo garantizar que la información se maneje de forma segura, especialmente si contiene datos sensibles.

```
Número_punt_crític|Codi_Districte|Nom_Districte|Codi_Barri|Nom_Barri|Nom_Carrer|Número|Latitud|Longitud|Motiu|Nombre_visites_febrer 2017|
Estat_febrer_2017|Nombre_visites_març_2017|Estat_març_2017|Nombre_visites_abril_2017|Estat_abril_2017|Nombre_visites_maig_2017|
Estat_maig_2017|Nombre_visites_juny_2017|Estat_juny_2017|Nombre_visites_juliol_2017|Estat_juliol_2017|Nombre_visites_agost_2017|
Estat_agost_2017|Districte d'alta: Mitjana edat dones|Districte d'alta: Mitjana edat homes|Districte de baixa: Mitjana edat dones|Districte d
baixa: Mitjana edat homes|població residente|Índex RFD BCN=100|posició
1|1|Ciutat Vella|1|el Raval|Passatge Elisabets|9|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|9.2|20|10.4|19|
14.7|21|15|21|22.3|31|32.1|32.4|32.9|47.274|74,6|48
2|1|Ciutat Vella|1|el Raval|Carrer de Valldoncella|49|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|11|16.5|21|6.9|19|
16.9|21|11.4|21|15.1|31|32.1|32.4|32.9|47.274|74,6|48
3|1|Ciutat Vella|1|el Raval|Carrer d'en Roig|1|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|11|7.7|21|9.8|20|11.5|20|11.8|21|9.6|31|32.1
32.4|32.9|47.274|74,6|48
4|1|Ciutat Vella|1|el Raval|Carrer Cardona|2|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|11.7|20|17.6|20|11.9|
21|13.8|21|9.3|31|32.1|32.4|32.9|47.274|74,6|48
5|1|Ciutat Vella|1|el Raval|Carrer Robadors |26|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|10|6.8|21|5.8|21|8.8|23|5.2|20|7.8|31|32.1|
32.4|32.9|47.274|74,6|48
6|1|Ciutat Vella|1|el Raval|Carrer de la Riera Baixa|26|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|11|11.9|21|10.6|20|13|21|10.3|21|
7.4|31|32.1|32.4|32.9|47.274|74,6|48
7|1|Ciutat Vella|1|el Raval|Plaça Hort de la Bomba|1|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|14|2.2|22|4.2|21|
6.5|21|4|21|3.6|31|32.1|32.4|32.9|47.274|74,6|48
8|1|Ciutat Vella|1|el Raval|Carrer de Sant Josep Oriol|13|41.4|2.2|Mal ús de la rec
21|4.3|21|2.6|21|1.7|31|32.1|32.4|32.9|47.274|74,6|48
```

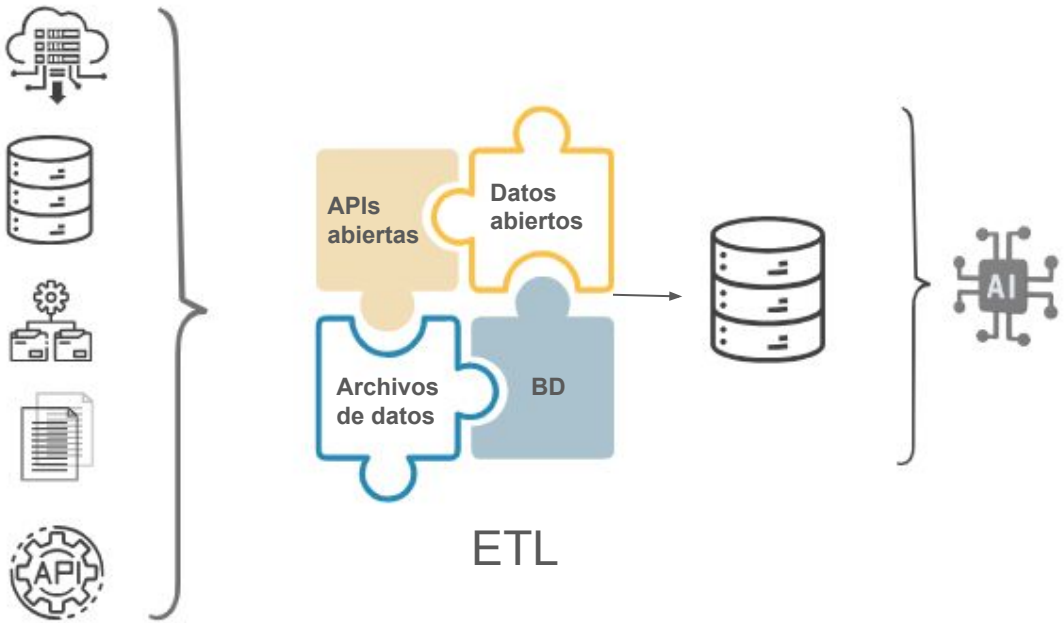
```
1,"4. Sant Pere, Santa Caterina i la Ribera","32.5","33.1",,"33.1","34.1"
2,"5. el Fort Pienc","32.6","33.9",,"33.3","34.9"
2,"6. la Sagrada Família","33.6",36,,"33.2","35.5"
2,"7. la Dreta de l'Eixample","36.2","39.6",,"34.3","36.7"
2,"8. l'Antiga Esquerra de l'Eixample","34.6","36.6",,"35.2","35.8"
2,"9. la Nova Esquerra de l'Eixample","34.1","35.3",,34,"34.6"
2,"10. Sant Antoni","34.5","34.9",,"33.5",35
```

```
▼<collection>
  ▼<record>
    <leader>01481nam a2200385 i 4500</leader>
    <controlfield tag="005">20210908155716.0</controlfield>
    <controlfield tag="007">ta</controlfield>
    <controlfield tag="008">210906s1584 it a r 000 0 ita d</controlfield>
    <controlfield tag="003">ES-GrU</controlfield>
    <controlfield tag="001">991014439313204990</controlfield>
  ▼<datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">UGR</subfield>
    </datafield>
  ▼<datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Cattaneo, Girolamo.</subfield>
    </datafield>
  ▼<datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Dell'arte militare libro quinto, de gli auuertimenti, et essamini intorno a quelle cose che richiede a
      vn bombardiero, cosi circa all'Artiglieria, come anco a fuochi arteficiati / di M. Girolamo Cataneo nouarese</subfield>
    </datafield>
  ▼<datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">In Brescia :</subfield>
      <subfield code="b">appresso Thomaso Bozzola,</subfield>
      <subfield code="c">1584</subfield>
    </datafield>
  ▼<datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">[2], 39, [1] h. :</subfield>
      <subfield code="b">il. ;</subfield>
      <subfield code="c">4º</subfield>
    </datafield>
  ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Marca tip. en port.</subfield>
    </datafield>
  ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Colofón</subfield>
    </datafield>
```

Integración de datos

Se entiende por integración de datos el conjunto de aplicaciones, productos, técnicas y tecnologías que permiten una **visión única consistente** de los datos

(Josep Curto Díaz)



¿Por qué es importante la integración de datos?

Desempeña un **papel fundamental**, los datos pueden estar almacenados en diversas fuentes:

- plataformas internas: sistema de información propio, sistemas operacionales y decisionales, cloud privado
- externas: IoT, cloud público, dispositivos móviles, datos abiertos, sensores

La información de cada sistema es valiosa, pero su **potencial** emerge **cuando se combinan.**

- facilita un análisis completo al tener visión total de los datos
- mejora la toma de decisiones en base a datos completos y fiables, no en información fragmentada
- reduce la duplicidad y mantiene la consistencia

Técnicas de integración de datos

- Propagación de datos
- Consolidación de datos
- Federación de datos
- CDC (Change Data Capture)
- Técnicas híbridas

Técnicas de integración de datos

Propagación de datos

Se copian los datos de la fuente al destino: **sincronización de datos**.

Los cambios se propagan de forma **síncrona o asíncrona**.

Se puede **activar por eventos o actualizaciones** específicas en el sistema origen.

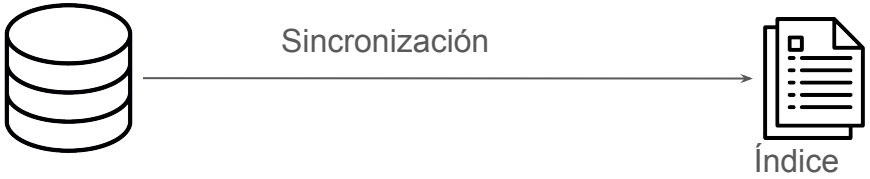
Busca asegurar la **consistencia** y facilitar los datos **en tiempo real** o con mínima latencia.

Útil cuando es necesario distribuir datos entre diferentes sistemas y ubicaciones.

Técnicas de integración de datos

Propagación de datos

1. Se realiza un cambio en la fuente.
2. **Detección**, se detecta un cambio. Se puede usar la técnica CDC (Change Data Capture) que rastrea los registros de transacciones o logs.
3. **Transmisión**, se transfiere solo el cambio detectado.
4. **Sincronización**, el sistema destino aplica el cambio, asegurando que el dato sea consistente con el origen.



APACHE LUCENE

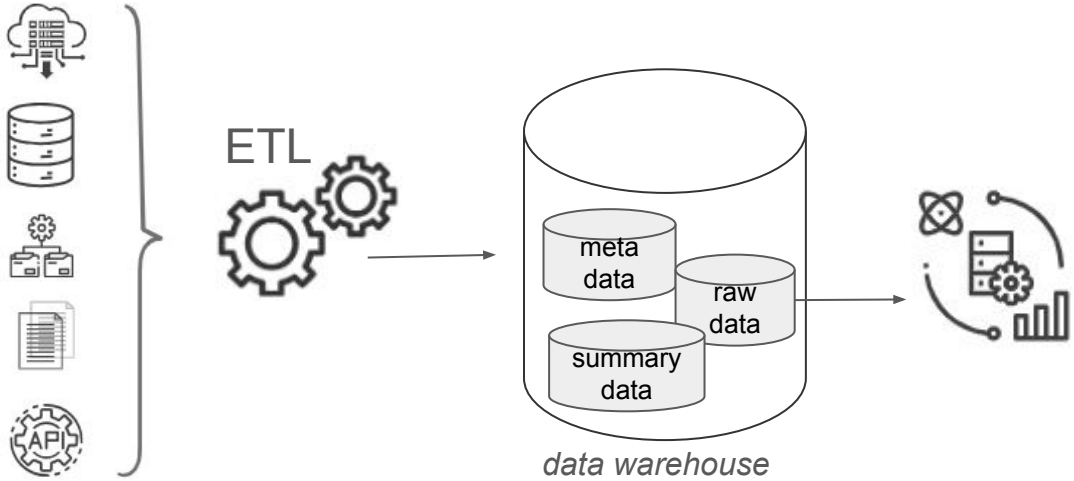
elasticsearch

Técnicas de integración de datos

Consolidación de datos

Agregación de datos de múltiples fuentes a un repositorio central como *data warehouse* o un *data lake*.

Se puede acceder a todos ellos de forma rápida y sencilla sin necesidad de consultar manualmente cada fuente origen.



Técnicas de integración de datos

Consolidación de datos

Puede existir cierta latencia en la actualización:

- **Requiere latencia baja**, se identifican los cambios en el origen para transmitir solo esos cambios, se suele usar algún tipo de técnica tipo **CDC** (Change Data Capture)
- **No requiere latencia baja**, se pueden usar procesos **batch** en intervalos prefijados

Técnicas de integración de datos

Federación de datos

Capa **virtual** que proporciona una **vista unificada** de los datos.

Facilita el **acceso y consulta en tiempo real** desde múltiples fuentes sin necesidad de moverlos ni replicarlos físicamente en un repositorio central.

Crea un **modelo lógico** con información de la estructura, su localización y define cómo se relacionan los datos de las distintas fuentes.

Optimiza la **fragmentación de la consulta original** y selecciona el camino de acceso a los datos más eficiente.

Técnicas de integración de datos

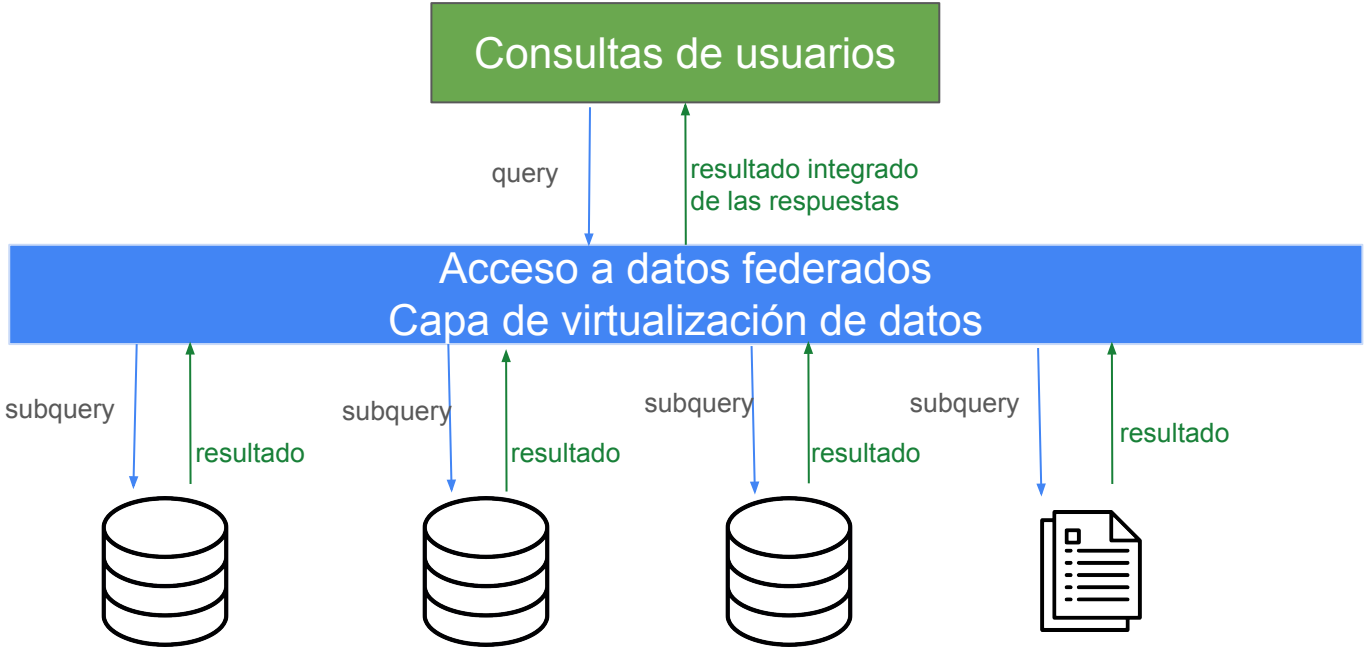
Federación de datos

Funcionamiento de una consulta:

1. La consulta se **descompone** en consultas individuales
2. Se **envía a cada una de las fuente** de datos involucradas (MySQL, Oracle, llamadas API…)
3. Los sistemas de origen **ejecutan** las consultas individuales
4. La capa **recopila** los respuestas
5. Las integran creando un **resultado único** que presenta al usuario

Técnicas de integración de datos

Federación de datos



Técnicas de integración de datos

Federación de datos

Ventajas:

- Proporciona **datos actualizados** porque accede directamente a la fuente.
- **Reduce costes de almacenamiento,** no requiere grandes servidores de almacenamiento (data warehouses o data lakes).
- Los **datos sensibles permanecen en su ubicación** original, facilitando el control de acceso y el cumplimiento normativo.
- Permite **integrar nuevas fuentes de datos de forma rápida** al crear el conector lógico, no requiere migración física.

Técnicas de integración de datos

CDC (*Change Data Capture*)

Identifica y captura cambios en el origen y los propaga a los entornos destino.

Mantiene sincronizados y consistentes los sistemas.

Técnicas de CDC:

- **Aplicación**: la aplicación genera el cambio y actualiza también en destino
- **Timestamp**: rastrea cambios en los datos de origen con timestamp
- **Triggers**: se activan tras una operación (INSERT, UPDATE, DELETE) y registran los detalles del cambio en una tabla de auditoría
- **Log**: se auditan los ficheros de logs en busca de cambios

Técnicas de integración de datos

CDC (*Change Data Capture*)

Técnica fundamental porque facilita la **sincronización con baja latencia** de los datos operativos a los sistemas de análisis.

Ventajas:

- Integración en **tiempo real**, los datos se actualizan inmediatamente
- Facilita la migración y replicación, **mantiene** origen y destino **sincronizado**
- **Eficiencia**, captura y mueve solo los cambios:
  - Reduce significativamente la sobrecarga en el sistema origen
  - Minimiza el uso de recursos de red y cómputo

Técnicas de integración de datos

Técnicas híbridas adaptadas al caso particular

En la práctica se suelen emplear varias técnicas de integración, lo que se denomina una técnica híbrida.

La técnica elegida dependerá de:

- los requisitos de negocio para la integración
- los requisitos tecnológicos y de las posibles restricciones presupuestarias

Tecnologías de integración de datos

Algunas de las más comunes:

- **ETL**: simplifican y automatizan el proceso de extracción, transformación y carga de datos
- **ELT**: los datos se cargan antes de cualquier transformación, ideales para casos donde es necesario volcar rápidamente grandes cantidades de datos
- **EII** (*Enterprise Information Integration*): basado en la federación de datos
- **EDR** (*Enterprise Data Replication*): detecta cambios en las fuentes de origen (CDC y propagación de datos)
- Integración basada en **APIs** (Application Programming Interfaces)

Tecnologías de integración de datos

ETL (*Extract Transform Load*)

ETL como una opción para la adquisición y preparación de los datos.

Extraen datos de las fuentes, transforman *raw data* siguiendo **reglas** de negocio para integración de datos, y los carga en su destino.

Documenta las transformaciones (logs, estadísticas asociadas a los procesos de movimiento de datos, etc).

- ETL basados en motor
- ETL de generación de código
- ETL integrado en la base de datos

Tecnologías de integración de datos

ETL (*Extract Transform Load*)

Se centran en:

- la **calidad:** la transformación de los datos garantiza que sean precisos y confiables
- la **organización** y **estructuración:** los datos deben estar bien organizados y listos para ser usados en su destino

Facilita:

- Adquisición y gestión de datos
- Extracción de datos
- Transformación de datos
- Gestión y administración de servicios
- Validación y auditoría: facilita auditar los pasos de transformación y garantizar la aplicación de las reglas de negocio de manera consistente

Tecnologías de integración de datos

ELT (*Extract Load Transform*)

Los datos se cargan antes de cualquier transformación.

- Prioriza la **velocidad** de la ingesta: útil cuando es necesario volcar rápidamente grandes cantidades de datos en un *data warehouse o data lake*.
- **Escalable**: aprovecha computación en la nube para las transformaciones.
- **Flexible**: los datos se almacenan en su formato original facilitando crear diferentes modelos de transformación sin tener que volver a extraer la información.
- Depende del coste y eficiencia del almacenamiento y computación en la nube.

Tecnologías de integración de datos

EII (*Enterprise Information Integration*)

Crea vista unificada de las distintos sistemas heterogéneos.

- Basada en la **federación de datos**.
- Proporcionar acceso en **tiempo real** y **consistente** a los datos de las distintas fuentes.
- Facilita la **interoperabilidad** y reduce la latencia, ya que los datos se ven tal como están en el momento de la consulta.
- No incurre en costes y complejidad asociados a la consolidación física como un data warehouse.

Tecnologías de integración de datos

EDR (*Enterprise Data Replication*)

Proceso de réplica de datos en grandes volúmenes a varias ubicaciones.

- Soporta las técnicas CDC (*Change Data Capture*) y propagación de datos.
- Utiliza la captura de datos modificados (CDC) para identificar los cambios incrementales y sincronizarlos con el destino.
- Asegura la disponibilidad y consistencia de los datos:
  - Consolida datos en tiempo real, regular o de forma esporádica
  - Mantiene copias de seguridad

Tecnologías de integración de datos

Integración basada en **APIs (Application Programming Interfaces)**

Proporciona una capa de abstracción entre diferentes sistemas.

Conexión de sistemas y aplicaciones a través de su API REST o SOAP.

- Facilita el intercambio de datos en tiempo real o con latencia mínima.
- Reduce las dependencias de los procesos ETL tradicionales en arquitecturas distribuidas.
- Google Cloud, Amazon Web Services, Microsoft Azure, etc.

Tecnologías de integración de datos

Uso de la integración de datos en diferentes tipologías de proyectos:

- Migración de datos
- Preparación de datos para modelos IA
- Procesos de calidad de datos
- BI (Business intelligence) y análisis de datos
- Data Warehousing

¿Cómo vamos a trabajar?

Flujos de trabajo

Existen diferentes herramientas que permiten trabajar con grandes volúmenes de datos y facilitan su análisis.

Para ello se diseñan y desarrollan flujos de trabajo que se pueden crear utilizando pasos o entradas unidas por saltos que pasan datos de un elemento al siguiente.

Los **flujos de trabajo** utilizan dos tipos de archivos:

1. Transformaciones que realizan tareas ETL.
2. Trabajos que organizan actividades de ETL, como la definición del flujo, dependencias y ejecución.

Pentaho Data Integration

Pentaho Data Integration (Kettle)

Entorno gráfico para diseño flujos de datos.

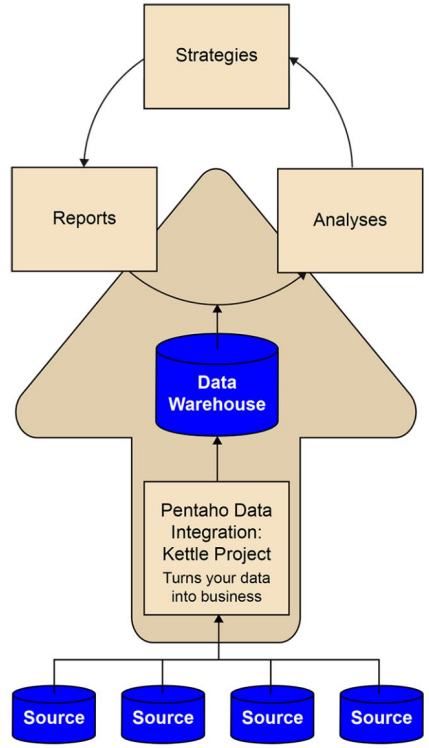
Servicio de integración de datos ETL (extract, transform, load).

Facilita el acceso a una amplia gama de fuentes de datos, incluidas bases de datos de código abierto y propietarias.

Arquitectura extensible mediante plugins y perspectivas

Algunos usos de Pentaho Data Integration

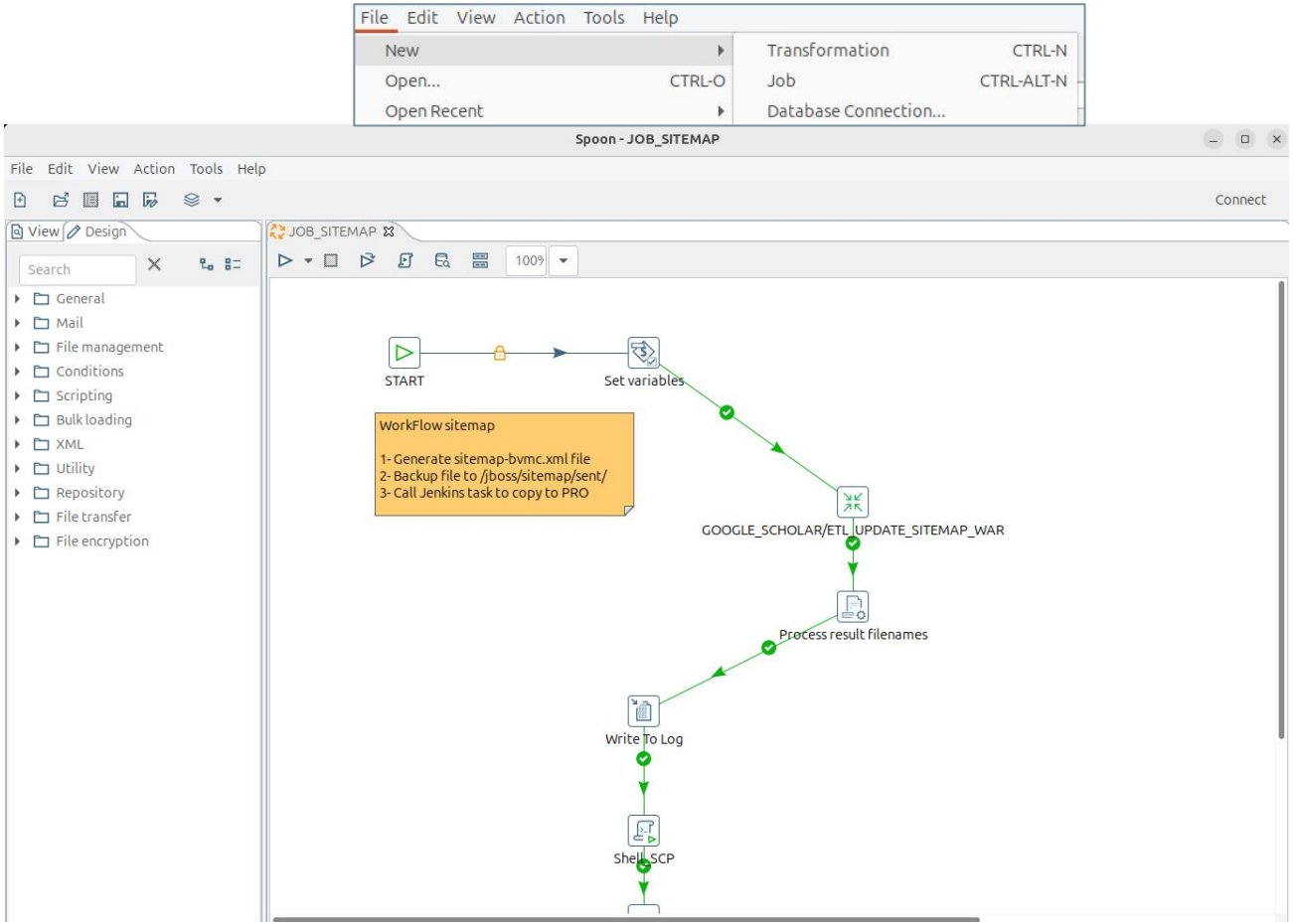
- Migración de datos entre diferentes sistemas
- Carga de grandes volúmenes de datos
- Limpieza de datos mediante transformaciones
- Integración de datos mediante herramientas ETL en tiempo real



Fuente: Pentaho.

Pentaho
Data Integration

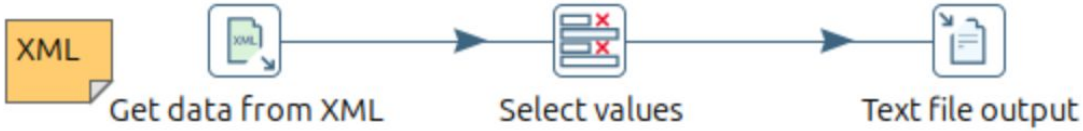
Entorno visual



Pentaho Data Integration

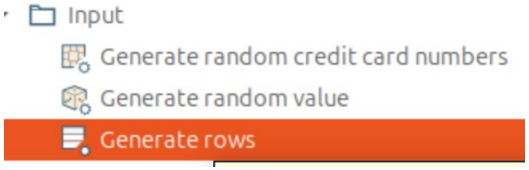
Transformations

- Proceso encargado de la **manipulación de datos** en el sentido más amplio (extracción, transformación y carga).
- Consta de uno o más **pasos** que realizan diferentes operaciones como leer, filtrar, limpiar o cargar datos en una base de datos.
- Los pasos están conectados mediante **saltos** que definen la dirección en el flujo de datos.
- La unidad de datos es la **fila**, y un **flujo de datos** es el movimiento de filas de un paso a otro paso.

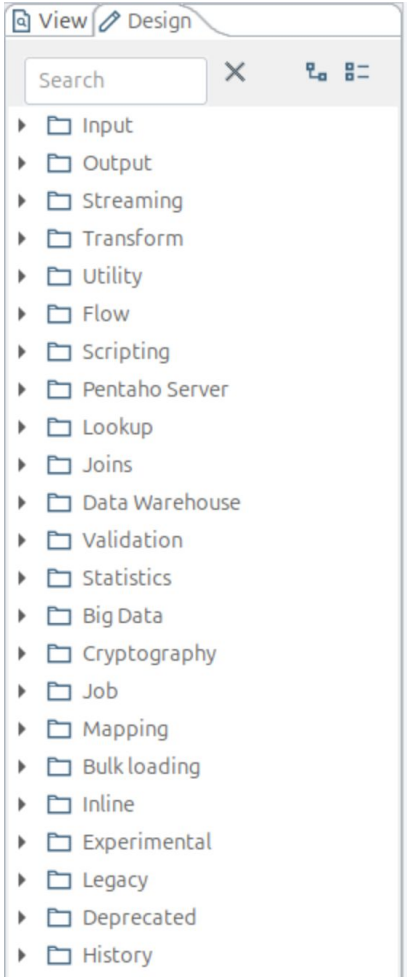


Pentaho Data Integration

Steps



- Cada paso una funcionalidad.
- Nombre único y amigable.
- Un paso lee y escribe filas de datos, excepción “Generate Rows” que solo escribe.
- Un paso puede escribir en uno o más flujos salientes:
  - Distribuir datos entre los diferentes flujos de salida (*round robin*)
  - Copiar datos, donde cada fila se envía a todos los flujos salientes
- Al ejecutar una transformación, se inician una o más copias de cada paso ejecutándose en su propio hilo.
- Durante la ejecución, todas las copias de pasos se ejecutan simultáneamente, con filas de datos fluyendo constantemente a través de sus saltos de conexión.



Pentaho Data Integration

Pasos para el diseño de *transformations*

Input, adquisición de datos desde diversas fuentes, BD, CSV, Excel, JSON, repositorios, etc

Output, carga datos en diferentes sistemas

Streaming, obtener datos de Kafka, JMS (Java Message Service) o MQTT (Standard for IoT Messaging)

Transform, ejecución de diversas operaciones con datos, como filtrar, ordenar, partir, añadir nuevos campos, mapear, etc

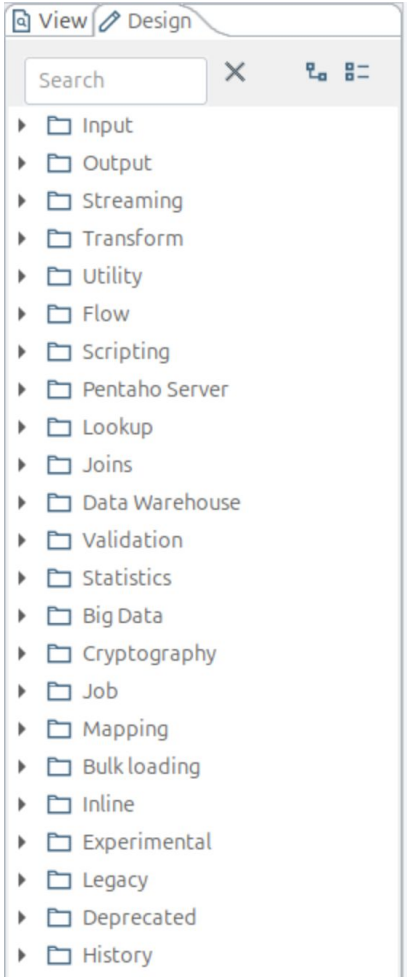
Utility, ejecución ssh commands, envío de mensajes syslog, etc

Flow, control de flujo de datos en las transformaciones

Scripting, SQLScript, JavaScript, expresiones regulares, fórmulas y expresiones java

Pentaho server, conexión al servidor de Pentaho para operaciones como la gestión de variables

Lookup, pasos para añadir información al flujo de datos mediante búsqueda en fuentes como BD, web services, HTTP, etc



Pentaho Data Integration

Pasos para el diseño de *transformations*

Joins, unión de filas en función de diferentes criterios

Data Warehouse, para trabajar con dimensiones

Validation, validaciones como tarjetas de crédito, datos, email o XSD

Statistics, operaciones estadísticas sobre un flujo de datos como group by

Big Data, cargar y extraer datos de Avro y MongoDB

Cryptography, uso de PGP (Pretty Good Privacy)

Job, realizar operaciones propias de un Job

Mapping, mapeo entre campos de entrada y salida

Bulk loading, cargas desde Vectores, MySQL, Oracle, PostgreSQL entre otras

Inline, embebido para operaciones con sockets

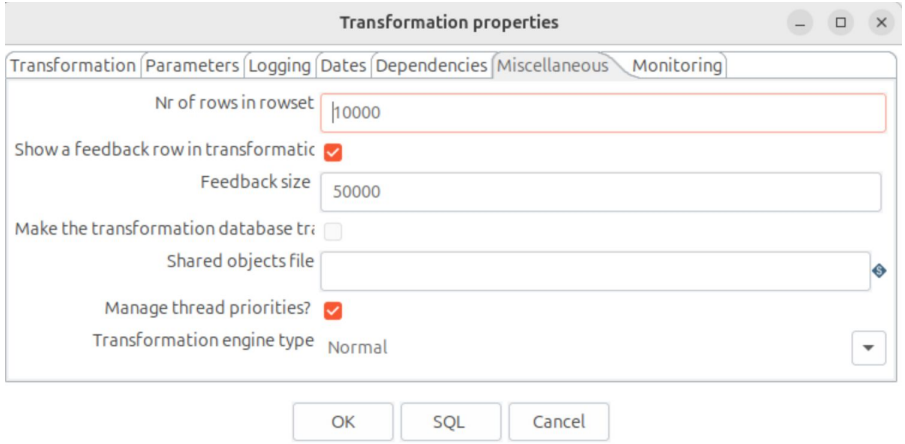
Experimental, pasos en fase de validación

Pentaho Data Integration

Transformation Hops

- Un salto define el sentido del flujo de datos entre los pasos.
- Representa un buffer de datos entre dos pasos llamado ***row set,*** configurable en *transformation settings*.

  Cuando el row set está lleno, el paso que escribe se detiene (no pasa datos) y el paso que realiza la lectura esperará hasta que tenga filas disponibles de nuevo.



- No se permiten bucles en transformaciones.

Pentaho Data Integration

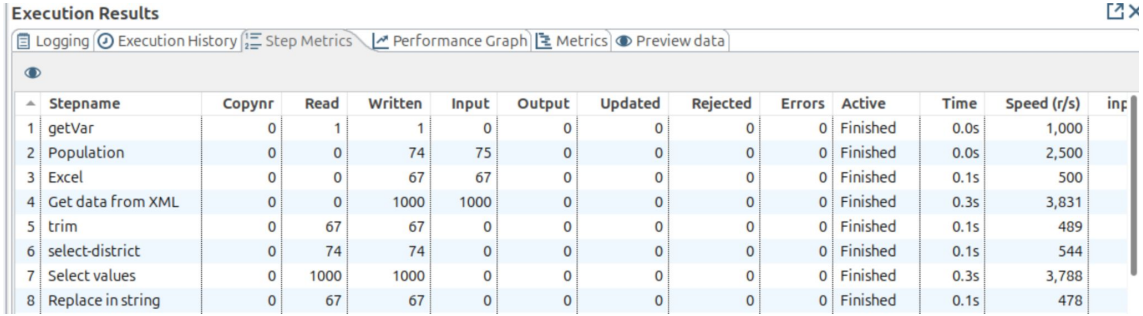
Parallelism

- Ejecución en paralelo en hilos separados.
- Permite procesamiento de datos en streaming con un consumo mínimo de memoria.
- No es posible definir un orden de ejecución.
- No es necesario identificar el inicio o fin de la transformación

Todos se ejecutan en paralelo

Se ejecutan mientras tengan datos de flujos entrantes

Se detiene cuando el último paso haya terminado



	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	inp
1	getVar	0	1	1	0	0	0	0	0	Finished	0.0s	1,000	
2	Population	0	0	74	75	0	0	0	0	Finished	0.0s	2,500	
3	Excel	0	0	67	67	0	0	0	0	Finished	0.1s	500	
4	Get data from XML	0	0	1000	1000	0	0	0	0	Finished	0.3s	3,831	
5	trim	0	67	67	0	0	0	0	0	Finished	0.1s	489	
6	select-district	0	74	74	0	0	0	0	0	Finished	0.1s	544	
7	Select values	0	1000	1000	0	0	0	0	0	Finished	0.3s	3,788	
8	Replace in string	0	67	67	0	0	0	0	0	Finished	0.1s	478	

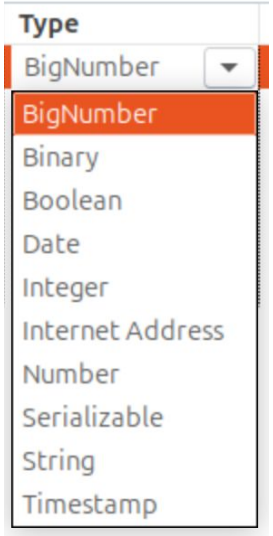
Pentaho Data Integration

Rows of Data

Los datos que pasan de un paso a otro durante el flujo de ejecución se presentan en forma ***row of data*.**

Colección de cero o más campos que pueden contener datos en cualquiera de los siguientes tipos:

- **String**, cualquier tipo de datos de carácter sin ningún límite.
- **Number**, número de punto flotante de doble precisión.
- **Integer**, entero largo con signo (64 bits).
- **BigNumber**, número con precisión arbitraria (ilimitada).
- **Date**, fecha y hora con precisión de milisegundos.
- **Boolean**, valor booleano puede contener verdadero o falso.
- **Binary**, los campos binarios pueden contener imágenes, sonidos, vídeos y otros tipos.
- **Internet address**
- **Serializable**
- **Timestamp**, para referencia temporal exacta



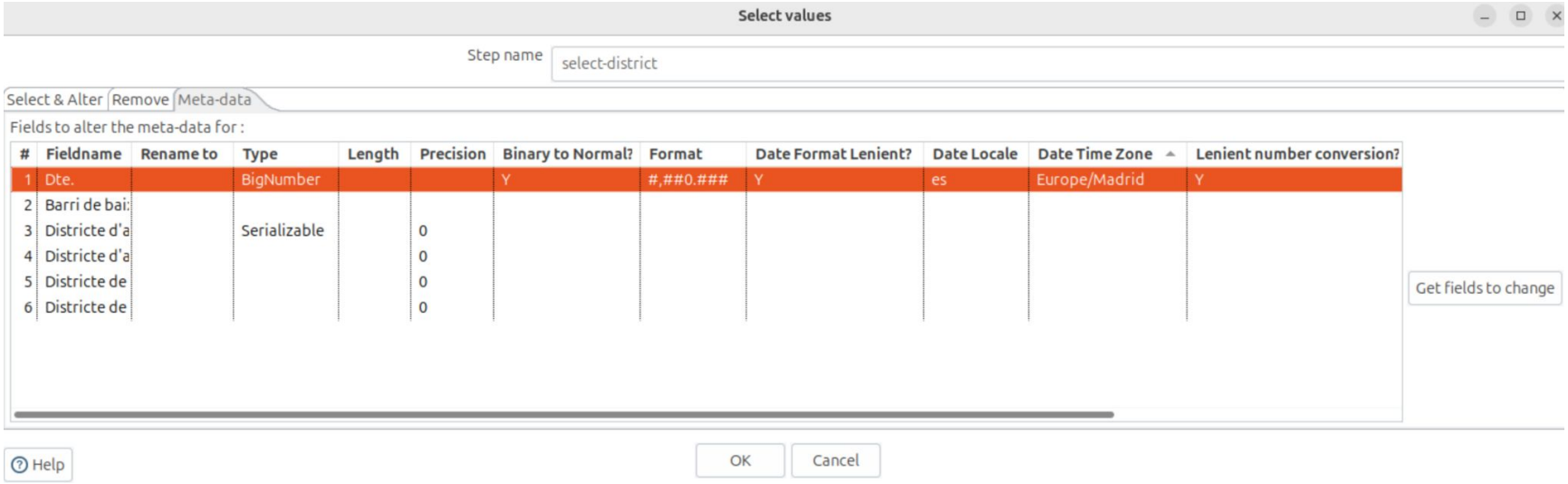
Pentaho Data Integration

Rows of Data

Cada paso permite describir las filas que se están publicando ***row metadata****:*

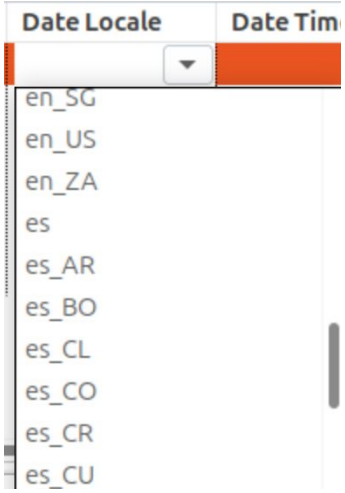
- **Name**, nombre del campo que debe ser único en una fila.
- **Data type**, tipo de dato del campo.
- **Length**, longitud de cadena o un número de un tipo de datos BigNumber.
- **Precision**, en decimal de un número BigNumber.
- **Format**, formato de representación (*conversion mask*).
    - Relevante en el caso de conversiones de tipo numérico (Número, Entero, BigNumber) o de Date a String.
    - Típico durante la vista previa de los datos en la interfaz o durante la serialización en texto o XML.
- **Date Format Lenient?** Y/N conversión de datos flexible permite flexibilidad y tolerancia a datos no estándar o mal formado durante la conversión.
- **Date Locale**, selección del idioma y dialecto.
- **Date Time zone**, selección de *time zone*.
- **Lenient number conversion?** Y/N , conversión flexible.
- **Encoding**, selección de la codificación.
- **Decimal**: El símbolo decimal de un número, este está definido culturalmente y suele ser punto (.) o una coma (,).
- **Grouping**: El símbolo de agrupación, esté está definido culturalmente y suele ser un coma (,), un punto (.), o una comilla simple (').
- **Currency,**

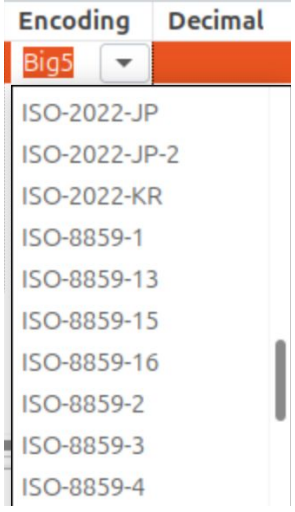
Pentaho Data Integration



#	Fieldname	Rename to	Type	Length	Precision	Binary to Normal?	Format	Date Format Lenient?	Date Locale	Date Time Zone	Lenient number conversion?
1	Dte.		BigNumber			Y	#,##0.###	Y	es	Europe/Madrid	Y
2	Barri de bai:										
3	Districte d'a		Serializable		0						
4	Districte d'a				0						
5	Districte de				0						
6	Districte de				0						

Pentaho Data Integration





Pentaho Data Integration

A tener en cuenta en el diseño de ETLs

- Todas las filas deben tener el mismo diseño o estructura.
  - Importante en el caso de flujos de datos distintos que van al mismo paso destino, el diseño de las filas de cada uno de estos flujos debe tener los mismos campos, con los mismos tipos de datos y en el mismo orden.
- Más allá del tipo de datos y el nombre, los metadatos del campo no se aplican durante la ejecución de una transformación (disponible solo en algunos pasos).
- De forma predeterminada, las cadenas vacías "" se consideran equivalentes a NULL.
- El comportamiento de que las cadenas vacías y NULL se consideren equivalentes pero se puede cambiar, modificando la variable KETTLE_EMPTY_STRING_DIFFERS_FROM_NULL → NULL vs Empty String. If this setting is set to Y, an empty string and null are different. Otherwise they are not.

Pentaho Data Integration

Variables PDI que marcan el funcionamiento de los tipos de datos

Kettle properties

Enter the values for the kettle.properties file

	Variable name	Value	Description
23	KETTLE_CORE_JOBENTRIES_FILE		The name of the environment variable that will contain the alternative location of the kettle-job-entries.xml file.
24	KETTLE_CORE_STEPS_FILE		The name of the environment variable that will contain the alternative location of the kettle-steps.xml file. You can use this to customize the list of available internal steps outside of the codebase.
25	KETTLE_DATA_REFINERY_HTTP_CLIENT_TIMEOUT	2000	A variable in ms to configure timeout inner http client data refinery plugin.
26	KETTLE_DATEDIFF_DST_AWARE	N	This environment variable is used to define how DateDiff ('Modified JavaScript value' step) performs regarding DST (Daylight Savings Time): "N" (default) - uses local time, meaning it assumes a continuum between the two given dates ignoring any time-hop that may had occurred between them "Y" - uses UTC, meaning it will take into acount any time-hop that may had occurred between the two given dates
27	KETTLE_DEFAULT_BIGNUMBER_FORMAT		The name of the variable containing an alternative default bignumber format
28	KETTLE_DEFAULT_CHECKSUM_EVALUATION_METHOD	BYTES	This environment variable is used to define which calculation method is to be used by the 'Add a Checksum' step: "BYTES" (default) - based on Byte representation of fields; as in versions since 8.1 "PENTAHO_STRINGS" - based on Pentaho String representation of fields (applying format masks); as in versions until 7.1 "NATIVE_STRINGS" - based on Native String representation of fields; as in version 8.0
29	KETTLE_DEFAULT_DATE_FORMAT		The name of the variable containing an alternative default date format
30	KETTLE_DEFAULT_INTEGER_FORMAT		The name of the variable containing an alternative default integer format
31	KETTLE_DEFAULT_NUMBER_FORMAT		The name of the variable containing an alternative default number format
32	KETTLE_DEFAULT_SERVLET_ENCODING		Defines the default encoding for servlets, leave it empty to use Java default encoding

OK Cancel

Pentaho Data Integration

Data Conversion

La conversión de datos se lleva a cabo explícitamente en el paso “*Select values*”.

Se puede cambiar el tipo de datos de un campo, o implícitamente, por ejemplo cuando se almacenan datos numéricos en un columna `varchar` en una base de datos relacional. Ambos tipos de conversión de datos se gestionan exactamente de la misma manera utilizando una combinación de los datos y la descripción de los datos.

Pentaho Data Integration

Date to String Conversion

La representación de fecha interna contiene la información necesaria para representar cualquier fecha/hora con precisión.

Para la conversión entre el tipo de dato cadena y fecha, es necesario especificar el formato de conversión.

Conversion Mask (format)	**Result**
yyyy/MM/dd’T’HH:mm:ss.SSS	2009/12/06T21:06:54.321
h:mm a	9:06 PM
HH:mm:ss	21:06:54
M-d-yy	12-6-09

Pentaho Data Integration

Select values

Step name: select-district

Select & Alter | Remove | Meta-data

Fields to alter the meta-data for :

#	Fieldname	Rename to	Type	Length	P	E	Format	Date Format Lenient?	Date Locale	Date Time Zone	Lenient number con
1	Dte.		Date			Y	m:ss.SSS	Y	es	Europe/Madrid	Y
2	Barri de bai:										
3	Districte d'a		Serializable		0						
4	Districte d'a				0						
5	Districte de				0						
6	Districte de				0						

yyyy/MM/dd HH:mm:ss.SSS
yyyy/MM/dd HH:mm:ss.SSS XXX
yyyy/MM/dd HH:mm:ss
yyyy/MM/dd HH:mm:ss XXX
yyyyMMddHHmmss
yyyy/MM/dd
yyyy-MM-dd
yyyy-MM-dd HH:mm:ss
yyyy-MM-dd HH:mm:ss XXX
yyyyMMdd
MM/dd/yyyy
MM/dd/yyyy HH:mm:ss
MM-dd-yyyy
MM-dd-yyyy HH:mm:ss
MM/dd/yy
MM-dd-yy
dd/MM/yyyy
dd-MM-yyyy
yyyy-MM-dd'T'HH:mm:ss.SSSXXX
yyyy-MM-dd HH:mm:ss.SSS

Get fields to change

Help | OK | Cancel

Pentaho Data Integration

Numeric to String Conversion

Los datos numéricos (Number, Integer, BigNumber) se convierten en/desde cadenas utilizando estos componentes de la pestaña metadatos:

- Format
- Decimal symbol
- Grouping symbol
- Currency symbol

Value	Conversion Mask	Decimal Symbol	Grouping Symbol	Result
1234.5678	#,###.##	.	,	1,234.57
1234.5678	000,000.00000	,	.	001.234,56780
-1.9	#.00;-#.00	.	,	-1.9
1.9	#.00;-#.00	.	,	1.9
12	00000;-00000			00012

El formato para la conversión numérica determina cómo se representa ese valor en formato de texto.

No influye en la precisión real ni en el redondeo de los datos numéricos.

Se puede consultar todos los símbolos permitidos y las reglas de formato en la documentación del API de Java.

Pentaho Data Integration

Other Conversions

From	To	Description
Boolean	String	This is converted to Y or N unless the length is 3 or higher. In that case, the result is true or false.
String	Boolean	A case-insensitive comparison is made and Y, True, Yes, and 1 all convert to true. Any other String is converted to false.
Integer Date	Date Integer	The Integer long value is considered to be the number of milliseconds that passed since January 1, 1970, 00:00:00 GMT. For example, September 12th 2010 at noon is converted to Integer 1284112800000 and vice-versa.

Pentaho Data Integration

KETTLE_DEFAULT_BIGNUMBER_FORMAT	The name of the variable containing an alternative default bignumber format
KETTLE_DEFAULT_CHECKSUM_EVALUATION_METHOD	This environment variable is used to define which calculation method is to be used by the 'Add a Checksum' step: ● "BYTES" (default) - based on Byte representation of fields; as in versions since 8.1 ● "PENTAHO_STRINGS" - based on Pentaho String representation of fields (applying format masks); as in versions until 7.1 ● "NATIVE_STRINGS" - based on Native String representation of fields; as in version 8.0
KETTLE_TIMESTAMP_NUMBER_CONVERSION_MODE LEGACY	This environment variable is used to define how Timestamp should be converted to a number and vice-versa. Possible values: "LEGACY" (default), "MILLISECONDS", "NANOSECONDS".
KETTLE_DO_NOT_NORMALIZE_NULL_STRING_TO_EMPTY	Prevents Kettle from converting null strings to empty strings when KETTLE_EMPTY_STRING_DIFFERS_FROM_NULL is set to "Y".
KETTLE_EMPTY_STRING_DIFFERS_FROM_NULL .	value:N NULL vs Empty String. If this setting is set to Y, an empty string and null are different. Otherwise they are not
KETTLE_DO_NOT_NORMALIZE_SPACES_ONLY_STRING_TO_EMPTY,	Prevents Kettle from converting spaces only strings to empty string

Procesos ETL

Tema 5: Introducción a la integración de datos mediante ETLs. Parte II

Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Pentaho Data Integration

Pentaho suite Business Intelligence para los procesos de gestión de datos.

Open source multiplataforma creada en 2004, plataforma Java.

Disponible en dos ediciones: **Community Edition** (CE) y **Enterprise Edition** (EE).

Múltiples plugins creado por la comunidad de código abierto.

Pentaho Community wiki
https://pentaho-public.atlassian.net/wiki/spaces/COM/overview?mode=global

Pentaho Data Integration

Pentaho Data Integration (Kettle)

Entorno gráfico para diseño flujos de datos.

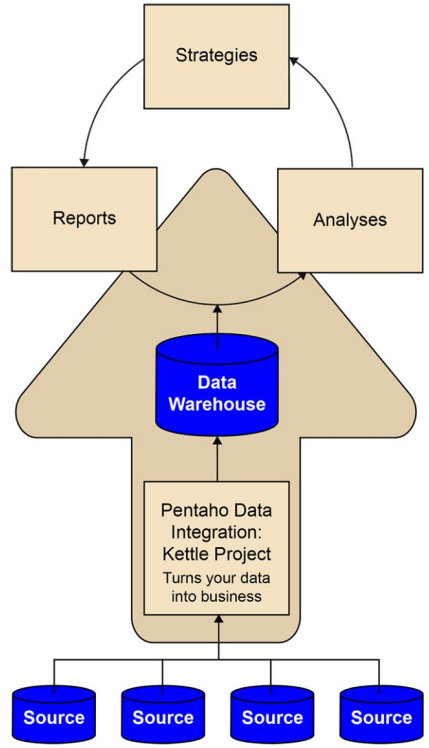
Servicio de integración de datos ETL (extract, transform, load).

Facilita el acceso a una amplia gama de fuentes de datos, incluidas bases de datos de código abierto y propietarias.

Arquitectura extensible mediante plugins y perspectivas

Algunos usos de Pentaho Data Integration

- Migración de datos entre diferentes sistemas
- Carga de grandes volúmenes de datos
- Limpieza de datos mediante transformaciones
- Integración de datos mediante herramientas ETL en tiempo real



Fuente: Pentaho.

Pentaho Data Integration

Componentes clave de Pentaho Data Integration

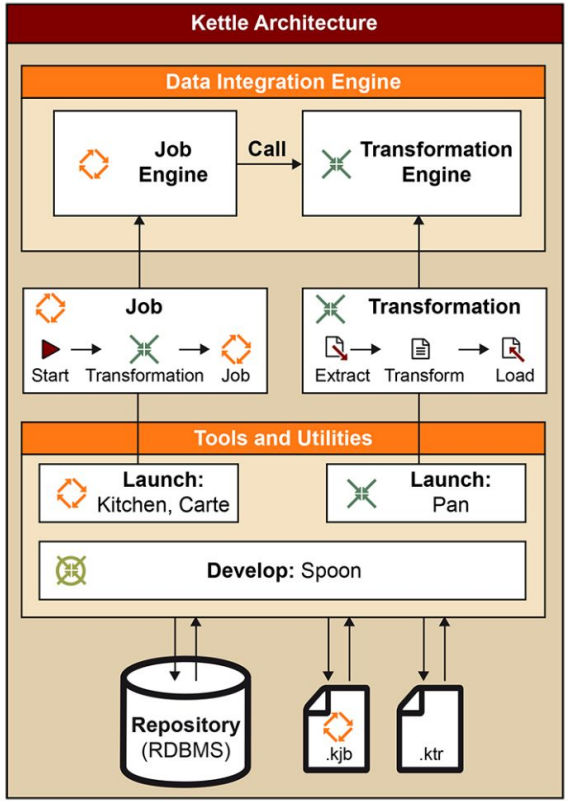
Spoon: entorno gráfico para creación de flujos de trabajo, transformaciones y trabajos.

- Realizar data warehouse
- Operaciones, como crear conexiones de datos, transformaciones o inserción de fórmulas, entre otras.
- Múltiples funcionalidades para realizar procesos de ejecución, transformación y carga de datos.

Pan: creación líneas de comandos para ejecutar transformaciones

Kitchen: ejecución de trabajos por línea de comandos

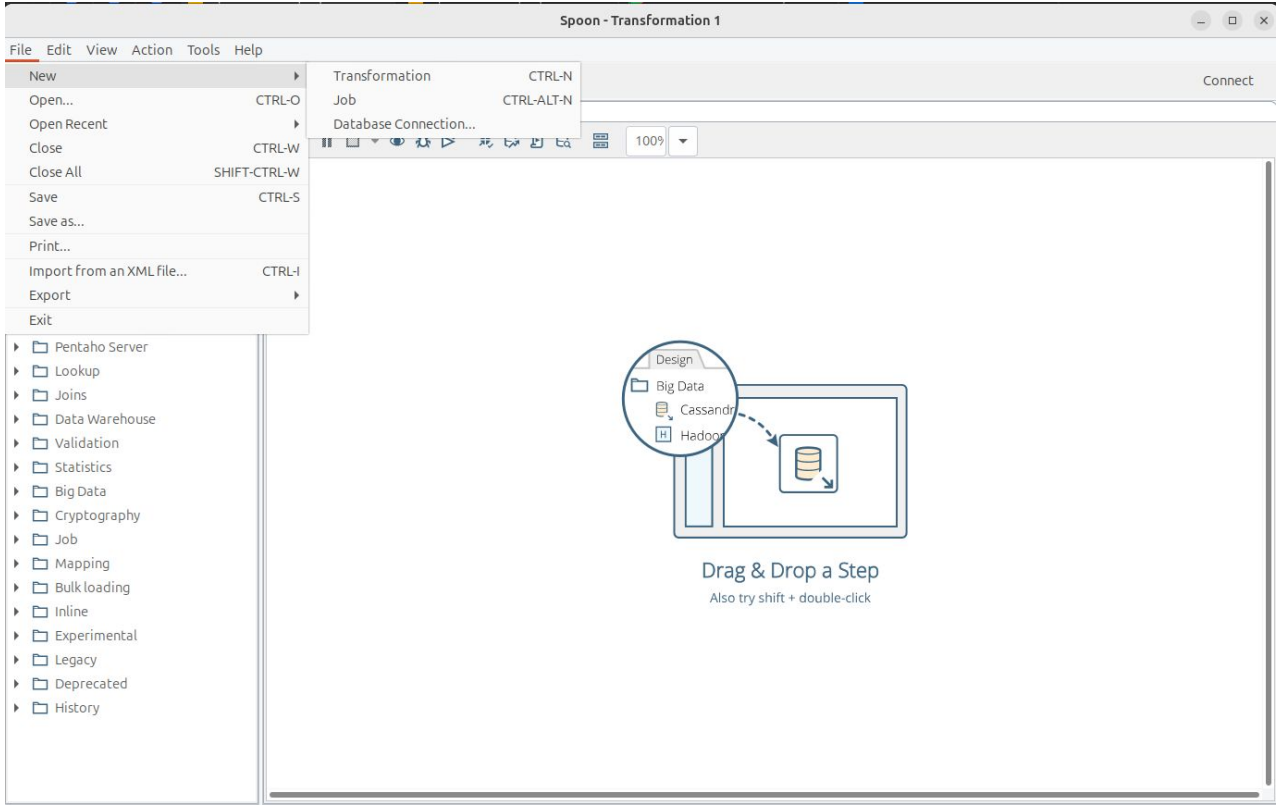
Carte: servidor web para ejecutar transformaciones y trabajos



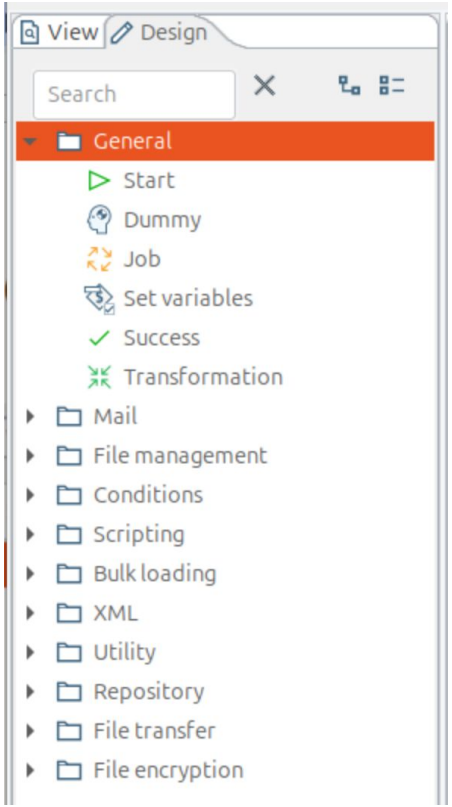
Fuente: Pentaho.

Pentaho
Data Integration

Entorno visual



Pentaho Data Integration



Pasos para el diseño de *jobs*

General, pasos de inicio de trabajo, ejecución transformaciones o trabajos entre otras operaciones

Mail, envío de correos, recuperación y validación de cuentas

File management, gestión de operaciones sobre ficheros y carpetas, HTTP, etc

Conditions, comprobaciones sobre conexiones, BD, ficheros, etc

Scripting, scripts en JavaScript, Shell y SQL

Bulk loading, cargas a MySQL y MSSQL

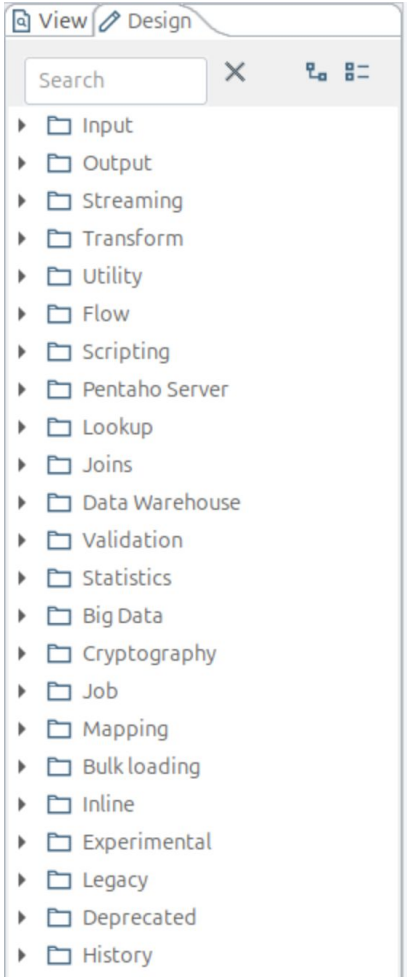
XML, validaciones XML, DTD, XSD y XSL

Utility, pasos para ejecución de transformaciones como ping, enviar información a syslogs, Nagios, etc

Repository, operaciones con el repositorio de transformaciones y trabajos.

File transfer, gestión ficheros FTP, FTPS y SFTP

File encryption : sistema PGP (Pretty Good Privacy) para envío y recepción de ficheros



Pentaho Data Integration

Pasos para el diseño de *transformations*

Input, adquisición de datos desde diversas fuentes, BD, CSV, Excel, JSON, repositorios, etc

Output, carga datos en diferentes sistemas

Streaming, obtener datos de Kafka, JMS (Java Message Service) o MQTT (Standard for IoT Messaging)

Transform, ejecución de diversas operaciones con datos, como filtrar, ordenar, partir, añadir nuevos campos, mapear, etc

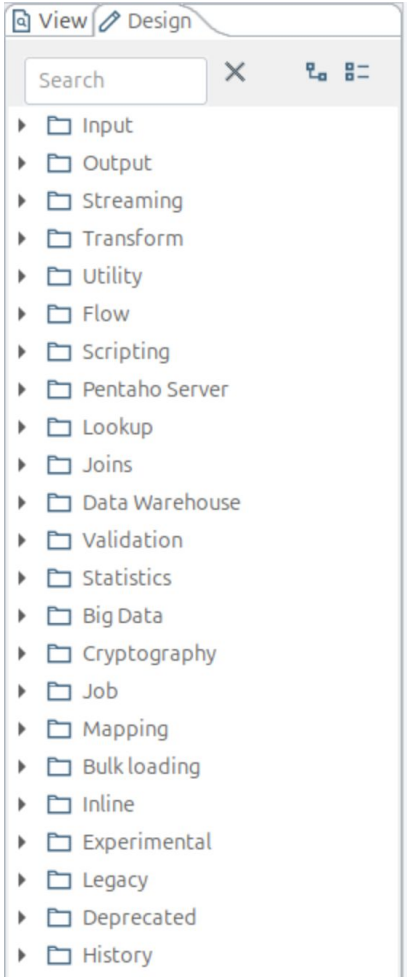
Utility, ejecución ssh commands, envío de mensajes syslog, etc

Flow, control de flujo de datos en las transformaciones

Scripting, SQLScript, JavaScript, expresiones regulares, fórmulas y expresiones java

Pentaho server, conexión al servidor de Pentaho para operaciones como la gestión de variables

Lookup, pasos para añadir información al flujo de datos mediante búsqueda en fuentes como BD, web services, HTTP, etc



Pentaho Data Integration

Pasos para el diseño de *Transformations*

Joins, unión de filas en función de diferentes criterios

Data Warehouse, para trabajar con dimensiones

Validation, validaciones como tarjetas de crédito, datos, email o XSD

Statistics, operaciones estadísticas sobre un flujo de datos como group by

Big Data, cargar y extraer datos de Avro y MongoDB

Cryptography, uso de PGP (Pretty Good Privacy)

Job, realizar operaciones propias de un Job

Mapping, mapeo entre campos de entrada y salida

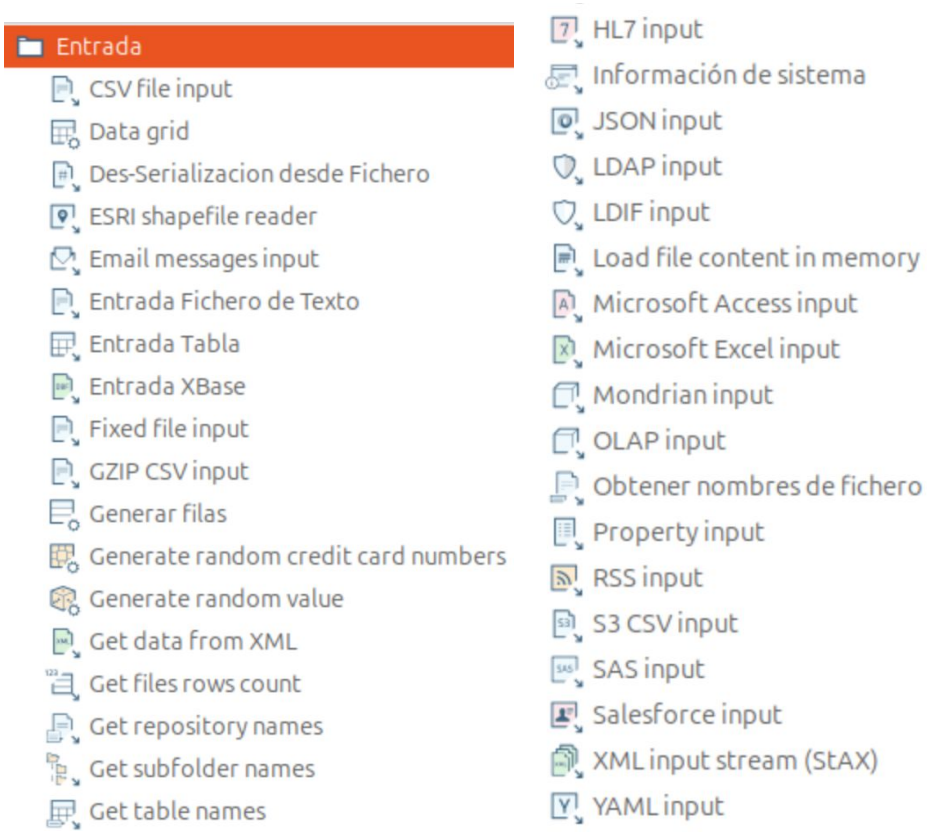
Bulk loading, cargas desde Vectores, MySQL, Oracle, PostgreSQL entre otras

Inline, embebido para operaciones con sockets

Experimental, pasos en fase de validación

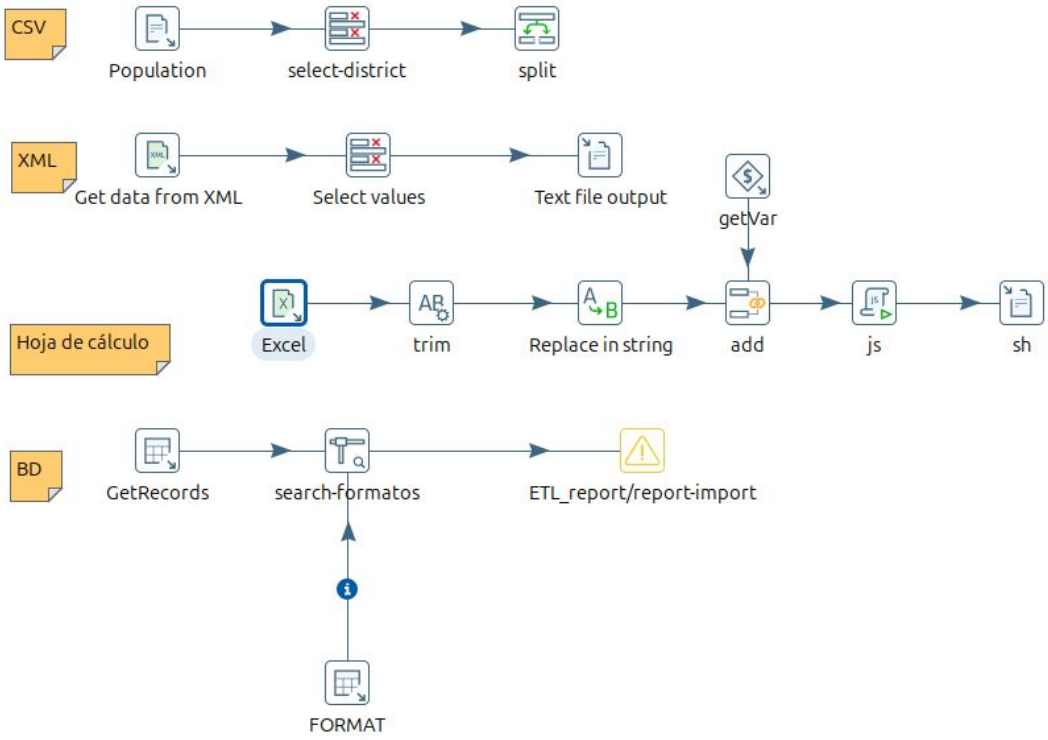
Pentaho Data Integration

Extract, adquisición de datos y distintos conectores para el acceso a las fuentes.



Pentaho Data Integration

Ejemplo 1: conexión a distintas fuentes de datos



```
Número_punt_crític|Codi_Districte|Nom_Districte|Codi_Barri|Nom_Barri|Nom_Carrer|Número|Latitud|Longitud|Motiu|Nombre_visites_febrer 2017|
Estat_febrer_2017|Nombre_visites_març_2017|Estat_març_2017|Nombre_visites_abril_2017|Estat_abril_2017|Nombre_visites_maig_2017|
Estat_maig_2017|Nombre_visites_juny_2017|Estat_juny_2017|Nombre_visites_juliol_2017|Estat_juliol_2017|Nombre_visites_agost_2017|
Estat_agost_2017|Districte d'alta: Mitjana edat dones|Districte d'alta: Mitjana edat homes|Districte de baixa: Mitjana edat dones|Districte d
baixa: Mitjana edat homes|població residente|Índex RFD BCN=100|posició
1|1|Ciutat Vella|1|el Raval|Passatge Elisabets|9|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|9.2|20|10.4|19|
14.7|21|15|21|22.3|31|32.1|32.4|32.9|47.274|74,6|48
2|1|Ciutat Vella|1|el Raval|Carrer de Valldoncella|49|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|11|16.5|21|6.9|19|
16.9|21|11.4|21|15.1|31|32.1|32.4|32.9|47.274|74,6|48
3|1|Ciutat Vella|1|el Raval|Carrer d'en Roig|1|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|11|7.7|21|9.8|20|11.5|20|11.8|21|9.6|31|32.1
32.4|32.9|47.274|74,6|48
4|1|Ciutat Vella|1|el Raval|Carrer Cardona|2|41.4|2.2|Mal ús de contenidors i/o papereres|0|No visitat|0|No visitat|11|11.7|20|17.6|20|11.9|
21|13.8|21|9.3|31|32.1|32.4|32.9|47.274|74,6|48
5|1|Ciutat Vella|1|el Raval|Carrer Robadors |26|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|10|6.8|21|5.8|21|8.8|23|5.2|20|7.8|31|32.1|
32.4|32.9|47.274|74,6|48
6|1|Ciutat Vella|1|el Raval|Carrer de la Riera Baixa|26|41.4|2.2|Efecte “crida”|0|No visitat|0|No visitat|11|11.9|21|10.6|20|13|21|10.3|21|
7.4|31|32.1|32.4|32.9|47.274|74,6|48
7|1|Ciutat Vella|1|el Raval|Plaça Hort de la Bomba|1|41.4|2.2|Mal ús de la recollida pneumàtica|0|No visitat|0|No visitat|14|2.2|22|4.2|21|
6.5|21|4|21|3.6|31|32.1|32.4|32.9|47.274|74,6|48
8|1|Ciutat Vella|1|el Raval|Carrer de Sant Josep Oriol|13|41.4|2.2|Mal ús de la rec
21|4.3|21|2.6|21|1.7|31|32.1|32.4|32.9|47.274|74,6|48
```

```
1,"4. Sant Pere, Santa Caterina i la Ribera","32.5","33.1",,"33.1","34.1"
2,"5. el Fort Pienc","32.6","33.9",,"33.3","34.9"
2,"6. la Sagrada Família","33.6",36,,"33.2","35.5"
2,"7. la Dreta de l'Eixample","36.2","39.6",,"34.3","36.7"
2,"8. l'Antiga Esquerra de l'Eixample","34.6","36.6",,"35.2","35.8"
2,"9. la Nova Esquerra de l'Eixample","34.1","35.3",,34,"34.6"
2,"10. Sant Antoni","34.5","34.9",,"33.5",35
```

```
▼<collection>
  ▼<record>
    <leader>01481nam a2200385 i 4500</leader>
    <controlfield tag="005">20210908155716.0</controlfield>
    <controlfield tag="007">ta</controlfield>
    <controlfield tag="008">210906s1584 it a r 000 0 ita d</controlfield>
    <controlfield tag="003">ES-GrU</controlfield>
    <controlfield tag="001">991014439313204990</controlfield>
  ▼<datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">UGR</subfield>
    </datafield>
  ▼<datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Cattaneo, Girolamo.</subfield>
    </datafield>
  ▼<datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Dell'arte militare libro quinto, de gli auuertimenti, et essamini intorno a quelle cose che richiede a
      vn bombardiero, cosi circa all'Artiglieria, come anco a fuochi arteficiati / di M. Girolamo Cataneo nouarese</subfield>
    </datafield>
  ▼<datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">In Brescia :</subfield>
      <subfield code="b">appresso Thomaso Bozzola,</subfield>
      <subfield code="c">1584</subfield>
    </datafield>
  ▼<datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">[2], 39, [1] h. :</subfield>
      <subfield code="b">il. ;</subfield>
      <subfield code="c">4º</subfield>
    </datafield>
  ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Marca tip. en port.</subfield>
    </datafield>
  ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Colofón</subfield>
    </datafield>
```

Extracción de datos

Extracción de datos

1. Extracción basada en bases de datos
2. Extracción basada en archivos
   a. VFS
   b. Trabajar con CSV/Excel
   c. Trabajar con archivos XML
3. Extracción basada en web
   a. Servicios web
   b. Cliente HTTP
   c. Extracción web basada en texto

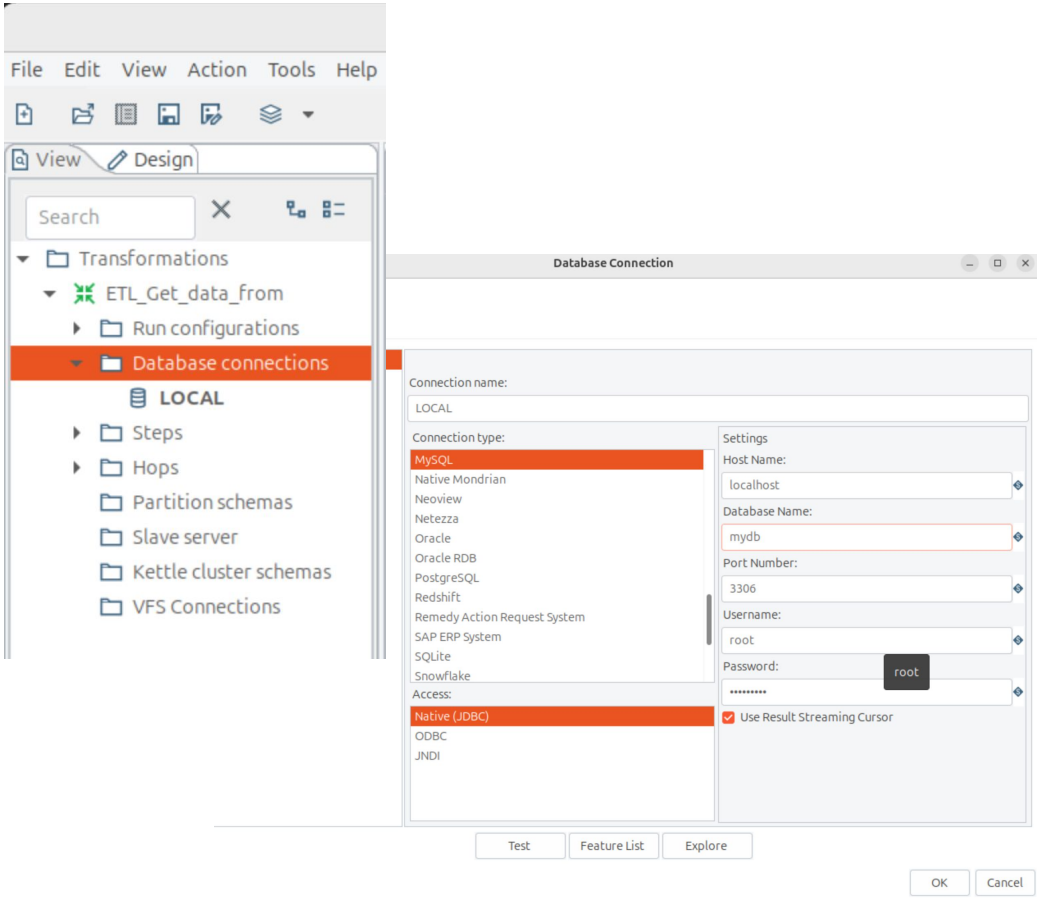
Extracción desde BD

Database Connection

Son utilizadas tanto por transformations como jobs.

La conexión con el RDBMS se establece en tiempo de ejecución.

Definir conexiones por sí solas no abre una conexión real a la base de datos.



Extracción desde BD

Database Connection

Database Connection

General
Advanced
Options
Pooling
Clustering

Parameters:

Parameter	Value
defaultFetchSize	500
useCursorFetch	true

Help...

General
Advanced
Options
Pooling
Clustering

Identifiers
- [x] Supports the boolean data type
- [x] Supports the timestamp data type
- [ ] Quote all in database
- [ ] Force all to lower case
- [ ] Force all to upper case
- [x] Preserve case of reserved words

The preferred schema name. Is used when no other schema is specified.

Test | Feature Li

General
Advanced
Options
Pooling
Clustering

Enable Connection Pooling

Pool Size — Initial: — Maximum:

Name	Value
defaultAutoCommit	true
defaultReadOnly	
defaultTransactionIsolation	
defaultCatalog	

Restore Defau

Test | Feature | Explore

General
Advanced
Options
Pooling
Clustering

Enable Clustering

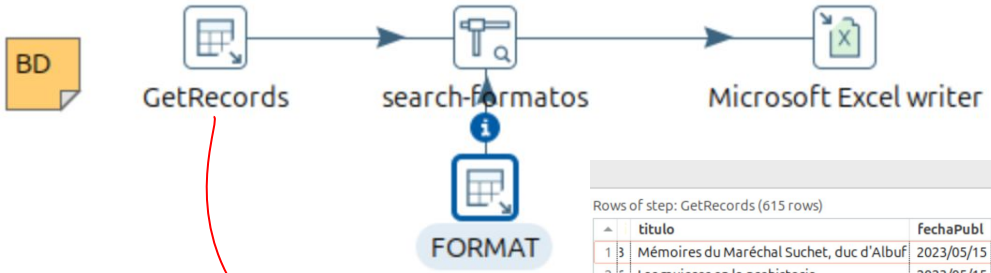
Parameters:

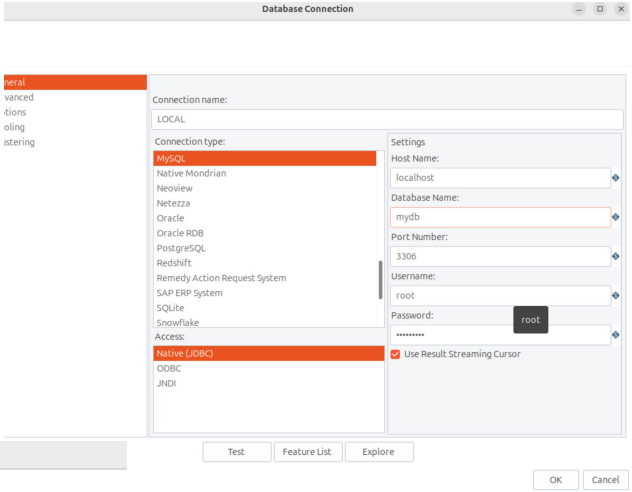
Partitior	Host Na	Port	Databas	User Nai	Passwor

Test | Feature | Explore

OK | Cancel

Extracción desde BD





Rows of step: GetRecords (615 rows)

		titulo	fechaPubl	lugarPublicacionOriginal
1	3	Mémoires du Maréchal Suchet, duc d'Albuf	2023/05/15	Anselin, 1834
2	5	Las mujeres en la prehistoria	2023/05/15	Museu de Prehistòria, 2008
3	9	Les dones en la prehistòria	2023/05/15	Diputació Provincial de Valèr
4	2	Libre de memories de diuersos successos e	2023/05/15	1600
5	5	[Còpia de documents relatius al Comte Dio	2023/05/15	1326
6	3	Otro título: Memorial y proyecto	2023/05/15	1769
7	1	Executoria del pleito seguido en el Consejo	2023/05/15	1795
8	4	Recopilacion verificada por Felipe Mateu, (	2023/05/15	1800
9	7	Dissertación historica sobre el Pontificado	2023/05/15	1740
10	0	Memoria histórica sobre el Origen y Visicitu	2023/05/15	1880
11	3	Carta de Sento Beseroles a Pascualo Valote	2023/05/15	1845
12	5	Carta de Bernat i Baldovi a D. Jose Brusca.	2023/05/15	1840
13	9	In Aristot. logicam commentarius...	2023/05/15	apud, Ioannem Bap. Marçal,
14	2	Academias morales de las musas...	2023/05/15	por Claudio Macè..., 1647
15	5	Segunda peregrinacion del dotor D. Pedro	2023/05/15	por Iayme de Bordazar, 1697
16	3	Glorias y excelencias del Gran Patriarca Sar	2023/05/15	1698
17	1	Las obras de Publio Virgilio Maron	2023/05/15	en la imprenta de Vicente Ca
18	4	Vida del hermano Iuan Bercmans flamenco	2023/05/15	por la viuda de Iuan Crisosto
19	7	De morborum et symptomatum differentii	2023/05/15	typis & expensis Michaelis So
20	0	Discursos predicables en las festiuidades de	2023/05/15	por Iuan Chrysostomo Garriz
21	3	El Pastor de Noche Buena : practica breve (	2023/05/15	por Silvestre Esparsa..., 1646

Extracción desde BD

Conexiones y transacciones

La conexión solo se utiliza durante la ejecución de un *job o transformation*.

En un job, cada entrada abre y cierra las conexiones de forma independiente.

En una transformación, debido a la naturaleza paralela, cada paso que utiliza una conexión, se abrirá una conexión a BD independiente e inicia una transacción independiente.

Generalmente ofrece excelente rendimiento para la mayoría de los casos, pero puede causar problemas de bloqueo e integridad referencial cuando se utilizan diferentes pasos de actualización en la misma tabla.

- Para evitar este problema se puede habilitar la opción "*Make transformation database transactional*", de forma que se establece una única conexión a esa BD.
- Se ejecuta un COMMIT cuando la transformación finaliza de manera correcta y un ROLLBACK en otro caso

Extracción basada en archivos

Virtual File Systems

La gestión flexible y uniforme de archivos es fundamental para cualquier herramienta ETL.

Especifica archivos como URL.

Utiliza backend VFS de Apache Commons que se encargará de la complejidad.

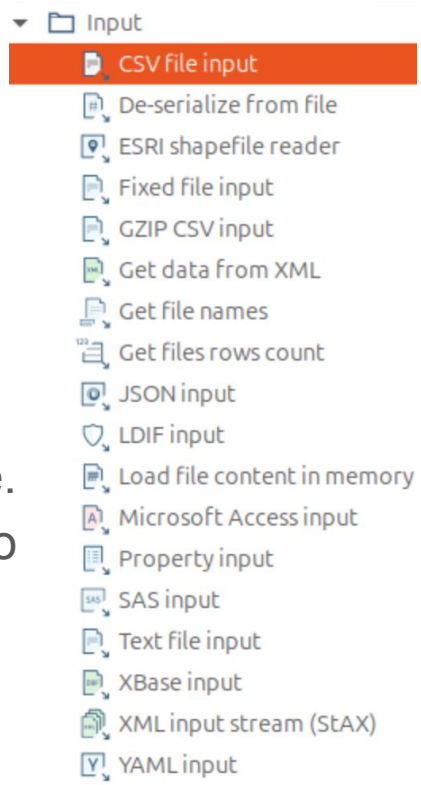
Por ejemplo, con Apache VFS, es posible procesar una selección de archivos dentro de un archivo .zip exactamente de la misma manera que procesaría una lista de archivos en una carpeta local. Para más información sobre cómo especificar archivos VFS,

Extracción basada en archivos

Nombre del archivo	Descripción
Nombre del archivo: /data/input /clientes.dat	Este archivo se define utilizando el método clásico (no VFS) y se encontrará y leer como tal.
Nombre del archivo: file:///data /input/clientes.dat	Este mismo archivo se leerá desde el sistema de archivos local mediante Apache Controlador VFS.
Trabajo: http://www. hervidor.be /GenerateRows.kjb	Este archivo se puede cargar en Spoon, ejecutar usando Kitchen y referenciar En la entrada de trabajo. Cada vez que se carga el archivo XML de forma transparente. desde el servidor web.
Carpeta: zip:file:///C: /entrada/datosdeventa.zip Comodín: .*\.txt$	Esta combinación de carpeta/comodín se puede ingresar en pasos como “Archivo de texto entrada.” La especificación del comodín buscará y leerá todos los archivos en el archivo zip especificado que termina en .txt.

Extracción basada en archivos

- Fáciles de gestionar con una herramienta ETL.
- Las operaciones R/W disponibles como pasos de transformación.
- La gestión de archivos está disponible en los pasos de *job* ***File management***: mover, copiar, crear, eliminar, comparar, comprimir y descomprimir.
- ***Text file output*** crea el archivo automáticamente si no existe.
- Tipos: ***Delimited*** (CSV recomendado) y ***Fixed*** (cada campo o columna tiene un ancho y una longitud designados).
- Importante configurar la codificación correcta, se puede seleccionar (UTF8 es más o menos el estándar).

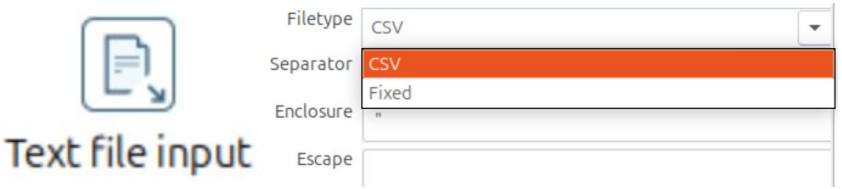


Extracción basada en archivos

La entrada más básica es ***CSV file input***. Permite gestionar un fichero y es necesario especificar el delimitador.

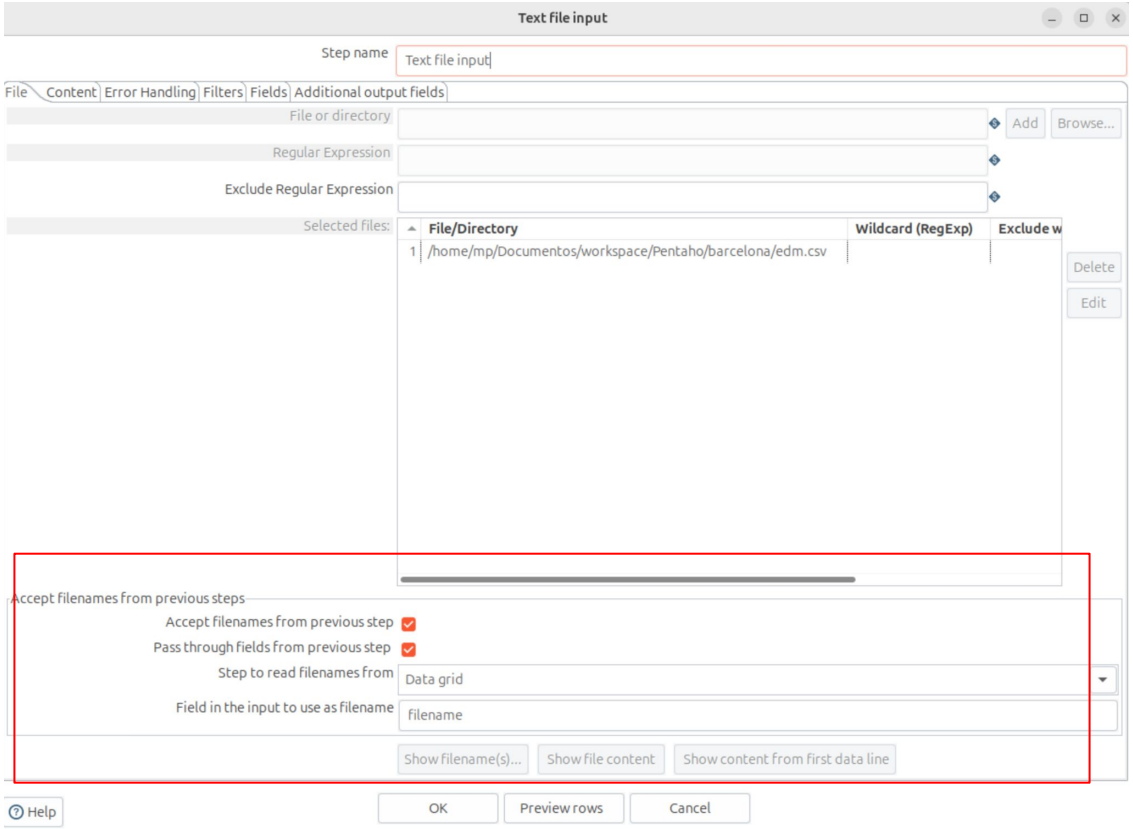
Text file input, versión más potente y completa de *CSV file input* y *Fixed file input:*

- Puede obtener el nombre del archivo del paso anterior
- Leer varios archivos en una sola ejecución
- Leer desde archivos comprimidos .zip o .gzip
- Mostrar el contenido del archivo de datos sin especificar la estructura
- Especificar carácter de escape
- Gestionar errores
- Filtrar
- Especificar formato de fecha local



Extracción basada en archivos

Text file input



Nombre del fichero lo recoge del paso anterior

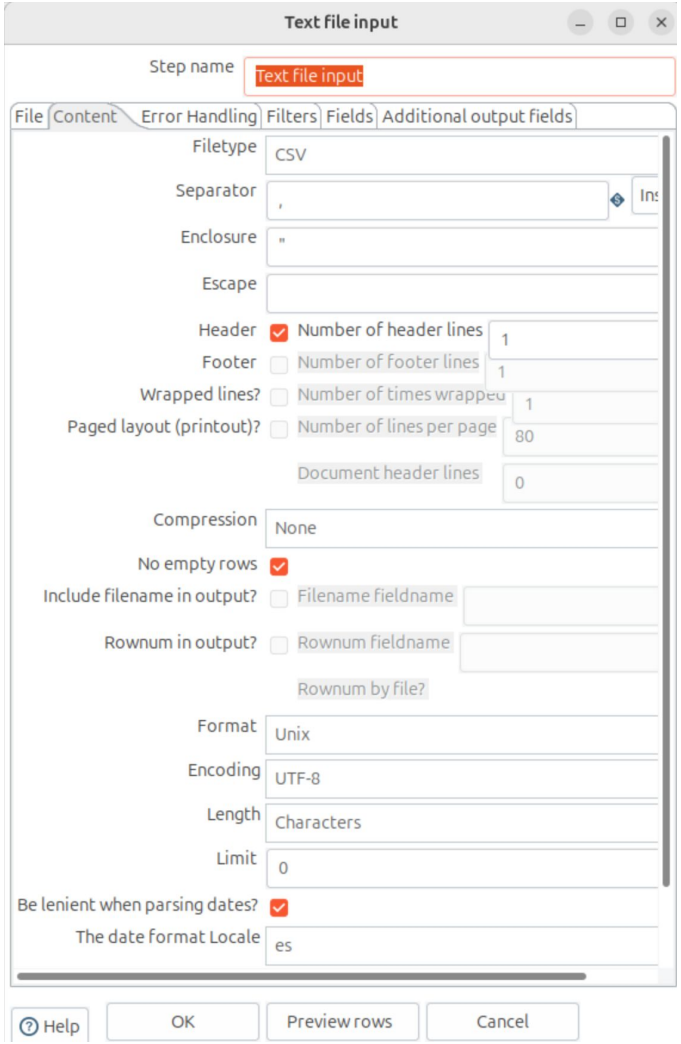
Indicamos el nombre del paso y el nombre del campo

Extracción basada en archivos

Text file input

Específica:

- Tipo de fichero: CSV
- Carácter de escape: ,
- Formato: Unix
- Encoding: UTF-8
- Especificar formato de fecha local: es



Extracción basada en archivos

Text file input

Selección de campos con los que trabajar de este archivo

Text file input

Step name: Text file input

File | Content | Error Handling | Filters | Fields | Additional output fields

	Name	Type	Format	Position	Length	Precision	Currency	Decimal	Group	Null if	Default	Trim type	Repeat
1	Dte.	String			0		$	.	,	-		none	N
2	Barri_de_baixa	String			9		$	.	,	-		none	N
3	Districte_d'alta:_Mitjana_edat_homes	String			48		$	.	,	-		none	N
4	Districte_d'alta:_Mitjana_edat_dones	BigNumber	#.#		4	1	$	.	,	-		none	N
5	Districte_de_baixa:_Mitjana_edat_homes	BigNumber	#.#		4	1	$	.	,	-		none	N
6	Districte_de_baixa:_Mitjana_edat_dones	String			0		$	.	,	-		none	N

Examine preview data

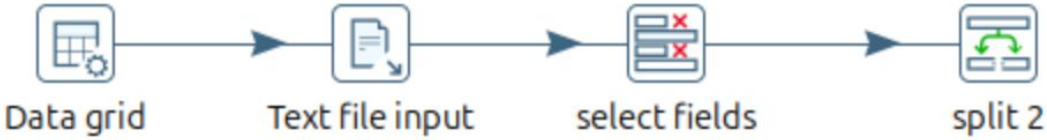
ows of step: Text file input (74 rows)

	filename	Dte.	Barri_de_baixa	Districte_d'
1	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	BARCELONA	<null>	33.2
2	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	1. el Raval	32.1
3	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	2. el Barri Gòtic	33.9
4	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	3. la Barceloneta	34.6
5	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	1	4. Sant Pere, Santa Caterina i la Ribera	32.5
6	/home/mp/Documentos/workspace/Pentaho/barcelona/edm.csv	2	5. el Fort Pienc	32.6

Extracción basada en archivos

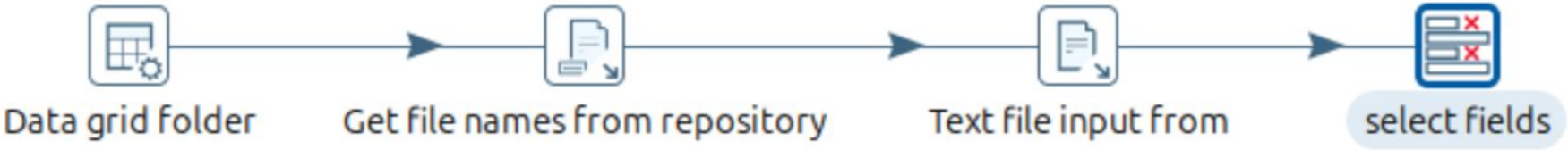
Uso del paso CSV input file
Particular de este tipo de fichero pero más limitado

Uso del paso Text input file
Más genérico y potente que los específicos CSV y Fixed



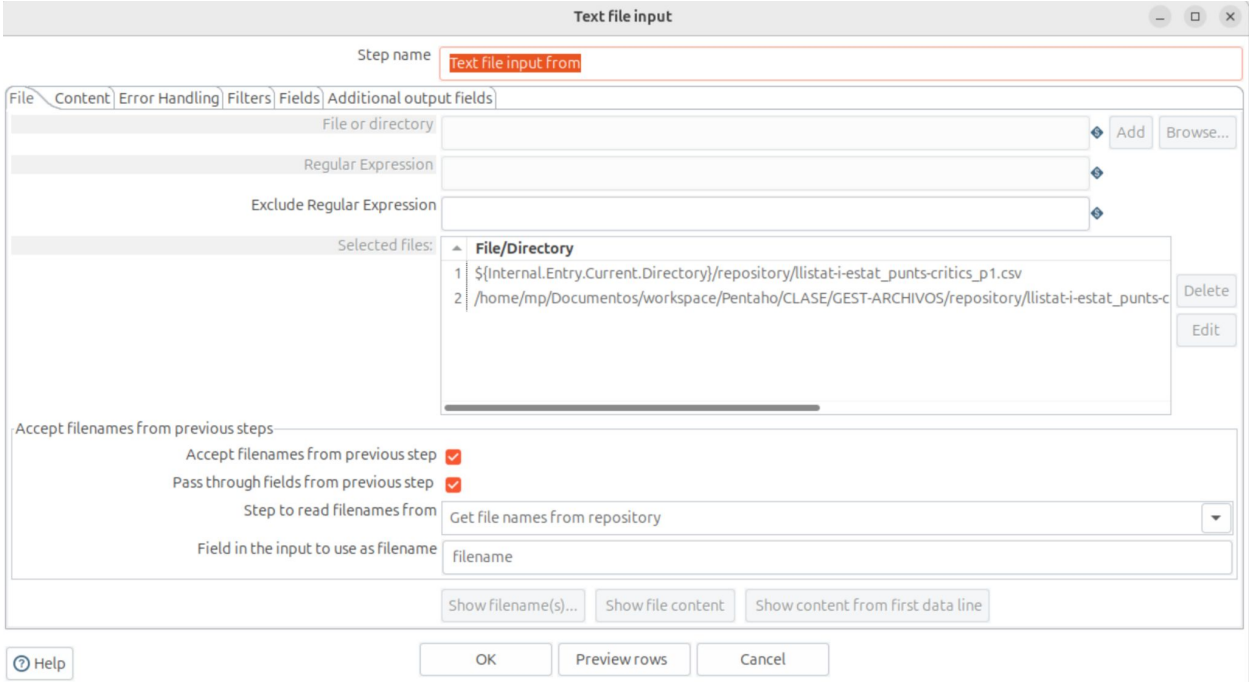
Extracción basada en archivos

Reading multiple files in a single run



Extracción basada en archivos

Reading multiple files in a single run



Extracción basada en archivos

Working with XML files

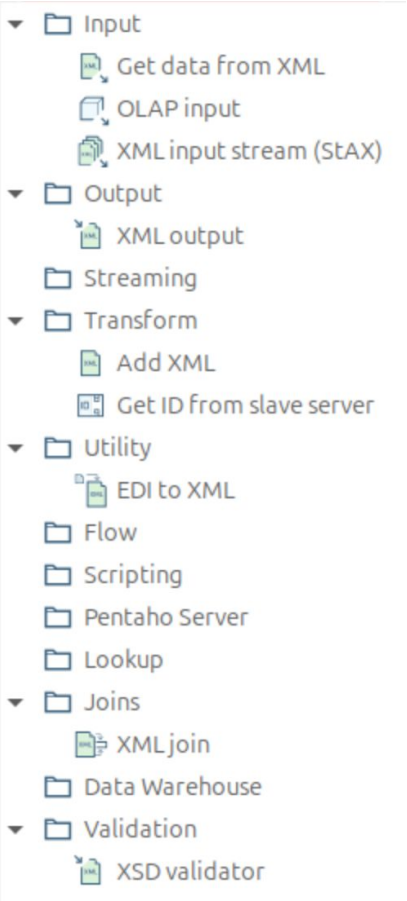
XML no es un formato simple y es necesario cumplir ciertas especificaciones, PDI facilita 4 opciones para validar la estructura:

- Check if XML file is well formed. Comprueba si las etiquetas de apertura y cierre están bien formadas. y la estructura anidada está bien equilibrada.
- DTD Validator. Verifica el contenido del archivo XML en base a un archivo de definición de tipo de datos, que puede ser interno (contenido dentro del archivo XML) o externo (archivo DTD separado).
- XSD Validator (Job). Verifica el contenido del archivo XML basándose en XML Schema
- Definition file
- XSD Validator (Transformation). Igual que el anterior, también puede verificar XML válido dentro de un campo de entrada específico, como una columna de base de datos que contiene datos XML

Extracción basada en archivos

Working with XML files

- Get data from XML
- XML input stream
- XML output
- XML Join
- XSD validator



Extracción basada en archivos

Working with XML files

Extracción basada en archivos

XML file

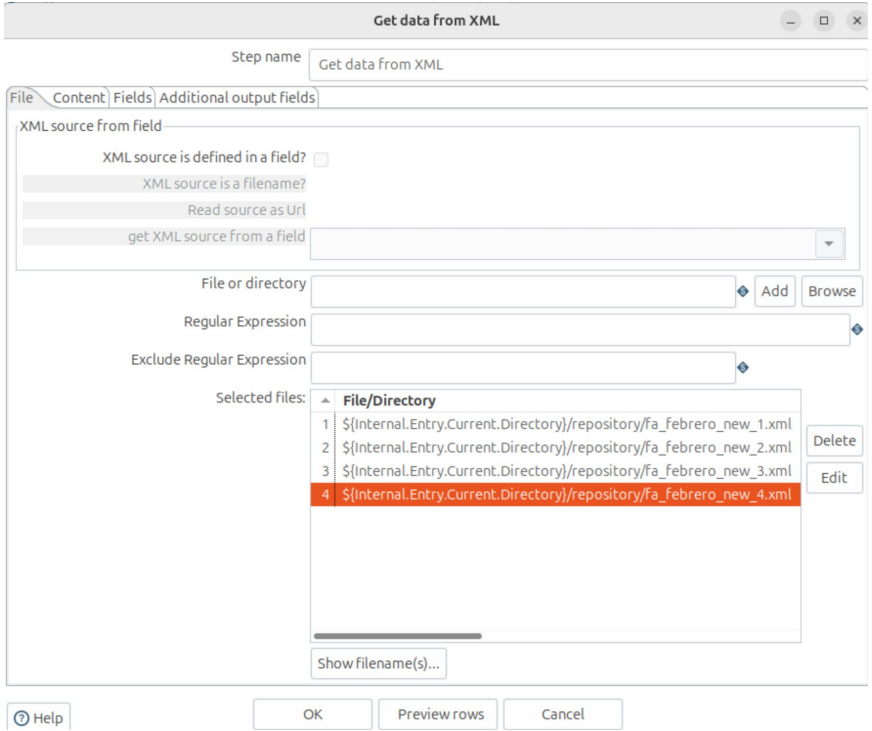
This XML file does not appear to have any style information associated with it. The document tree is shown below.

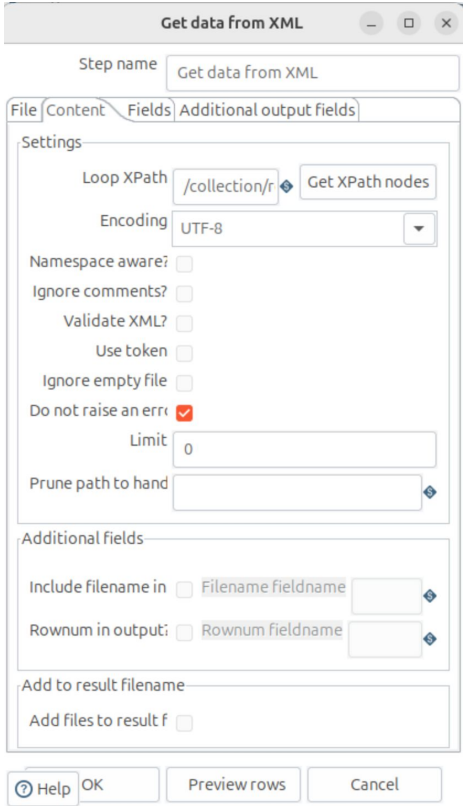

```
▼<collection>
  ▼<record>
    <leader>01481nam a2200385 i 4500</leader>
    <controlfield tag="005">20210908155716.0</controlfield>
    <controlfield tag="007">ta</controlfield>
    <controlfield tag="008">210906s1584 it a r 000 0 ita d</controlfield>
    <controlfield tag="003">ES-GrU</controlfield>
    <controlfield tag="001">991014439313204990</controlfield>
    ▼<datafield tag="040" ind1=" " ind2=" ">
      <subfield code="a">UGR</subfield>
    </datafield>
    ▼<datafield tag="100" ind1="1" ind2=" ">
      <subfield code="a">Cattaneo, Girolamo.</subfield>
    </datafield>
    ▼<datafield tag="245" ind1="1" ind2="0">
      <subfield code="a">Dell'arte militare libro quinto, de gli auuertimenti, et essamini intorno
      a quelle cose che richiede a vn bombardiero, cosi circa all'Artiglieria, come anco a fuochi
      arteficiati / di M. Girolamo Cataneo nouarese</subfield>
    </datafield>
    ▼<datafield tag="260" ind1=" " ind2=" ">
      <subfield code="a">In Brescia :</subfield>
      <subfield code="b">appresso Thomaso Bozzola,</subfield>
      <subfield code="c">1584</subfield>
    </datafield>
    ▼<datafield tag="300" ind1=" " ind2=" ">
      <subfield code="a">[2], 39, [1] h. :</subfield>
      <subfield code="b">il. ;</subfield>
      <subfield code="c">4º</subfield>
    </datafield>
    ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Marca tip. en port.</subfield>
    </datafield>
    ▼<datafield tag="500" ind1=" " ind2=" ">
      <subfield code="a">Colofón</subfield>
    </datafield>
```

Extracción basada en archivos

Working with XML files

Get data from XML





Extracción basada en archivos

Working with XML files

Get data from XML

Get data from XML

Step name: Get data from XML

File | Content | Fields | Additional output fields

	Name	XPath	Element	Result type	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type	Repeat
1	leader	leader	Node	Value of	String							none	N
2	nuevoID_001	controlfield[@tag=001]	Node	Value of	String							none	N
3	OCoLC_ID_035	datafield[@tag=035]/subfield[starts-with(., '(OCoLC)')]	Node	Value of	None							none	N
4	HANDLED_URL_856u	datafield[@tag=856]/subfield[@code='u']	Node	Value of	String							none	N
5	antiguoID_URL_962a	datafield[@tag=962]/subfield[@code='a']	Node	Value of	String							none	N
6	antiguoID_035	datafield[@tag=035]/subfield[starts-with(., '(ES-GrU)b')]	Node	Value of	String							none	N
7	antiguoID_907	datafield[@tag=907]/subfield[@code='a']	Node	Value of	String							none	N
8	nombre_245	datafield[@tag=245]/subfield[@code='a']	Node	Value of	String							none	N

Get fields

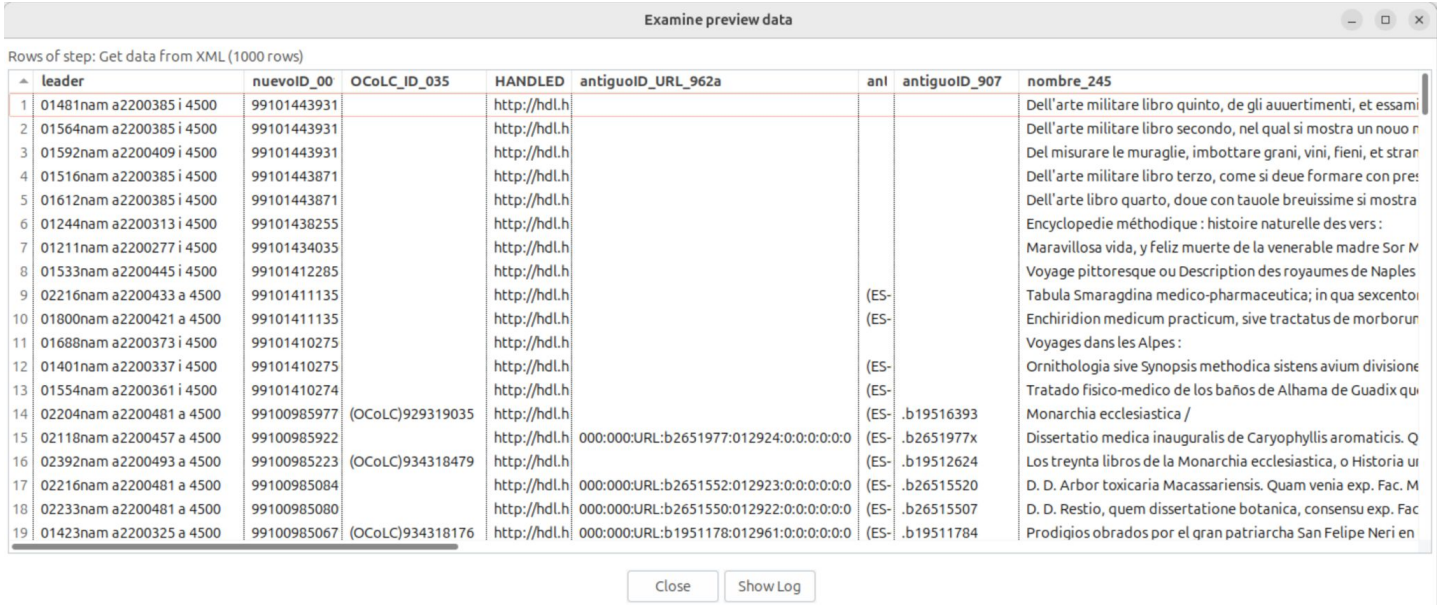
Help | OK | Preview rows | Cancel

Extracción basada en archivos

Working with XML files

Get data from XML

Preview rows

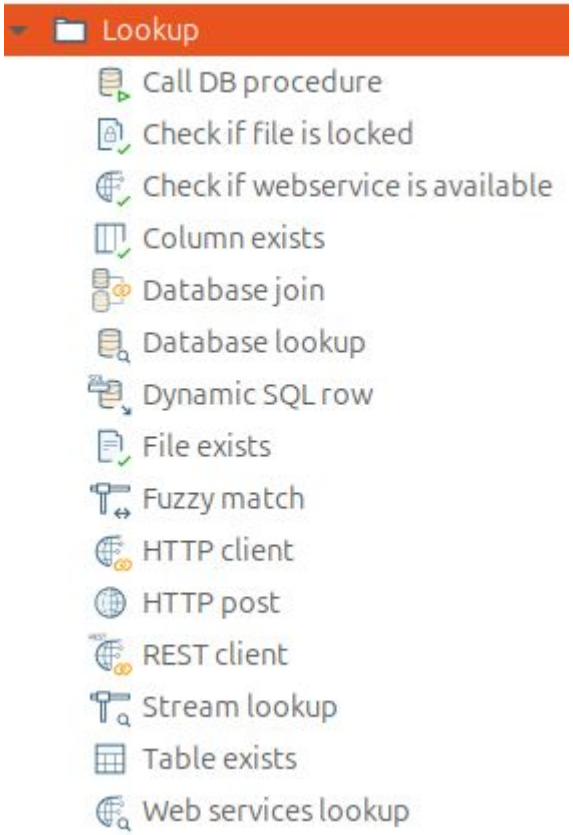


	leader	nuevoID_00	OCoLC_ID_035	HANDLED	antiguoID_URL_962a	anl	antiguoID_907	nombre_245
1	01481nam a2200385 i 4500	99101443931		http://hdl.h				Dell'arte militare libro quinto, de gli auuertimenti, et essami
2	01564nam a2200385 i 4500	99101443931		http://hdl.h				Dell'arte militare libro secondo, nel qual si mostra un nouo r
3	01592nam a2200409 i 4500	99101443931		http://hdl.h				Del misurare le muraglie, imbottare grani, vini, fieni, et stran
4	01516nam a2200385 i 4500	99101443871		http://hdl.h				Dell'arte militare libro terzo, come si deue formare con pres
5	01612nam a2200385 i 4500	99101443871		http://hdl.h				Dell'arte libro quarto, doue con tauole breuissime si mostra
6	01244nam a2200313 i 4500	99101438255		http://hdl.h				Encyclopedie méthodique : histoire naturelle des vers :
7	01211nam a2200277 i 4500	99101434035		http://hdl.h				Maravillosa vida, y feliz muerte de la venerable madre Sor M
8	01533nam a2200445 i 4500	99101412285		http://hdl.h				Voyage pittoresque ou Description des royaumes de Naples
9	02216nam a2200433 a 4500	99101411135		http://hdl.h		(ES-		Tabula Smaragdina medico-pharmaceutica; in qua sexcento
10	01800nam a2200421 a 4500	99101411135		http://hdl.h		(ES-		Enchiridion medicum practicum, sive tractatus de morborun
11	01688nam a2200373 i 4500	99101410275		http://hdl.h				Voyages dans les Alpes :
12	01401nam a2200337 i 4500	99101410275		http://hdl.h		(ES-		Ornithologia sive Synopsis methodica sistens avium divisione
13	01554nam a2200361 i 4500	99101410274		http://hdl.h		(ES-		Tratado fisico-medico de los baños de Alhama de Guadix qu
14	02204nam a2200481 a 4500	99100985977	(OCoLC)929319035	http://hdl.h		(ES-	.b19516393	Monarchia ecclesiastica /
15	02118nam a2200457 a 4500	99100985922		http://hdl.h	000:000:URL:b2651977:012924:0:0:0:0:0:0	(ES-	.b2651977x	Dissertatio medica inauguralis de Caryophyllis aromaticis. Q
16	02392nam a2200493 a 4500	99100985223	(OCoLC)934318479	http://hdl.h		(ES-	.b19512624	Los treynta libros de la Monarchia ecclesiastica, o Historia u
17	02216nam a2200481 a 4500	99100985084		http://hdl.h	000:000:URL:b2651552:012923:0:0:0:0:0:0	(ES-	.b26515520	D. D. Arbor toxicaria Macassariensis. Quam venia exp. Fac. M
18	02233nam a2200481 a 4500	99100985080		http://hdl.h	000:000:URL:b2651550:012922:0:0:0:0:0:0	(ES-	.b26515507	D. D. Restio, quem dissertatione botanica, consensu exp. Fac
19	01423nam a2200325 a 4500	99100985067	(OCoLC)934318176	http://hdl.h	000:000:URL:b1951178:012961:0:0:0:0:0:0	(ES-	.b19511784	Prodigios obrados por el gran patriarcha San Felipe Neri en

Extracción basada en web

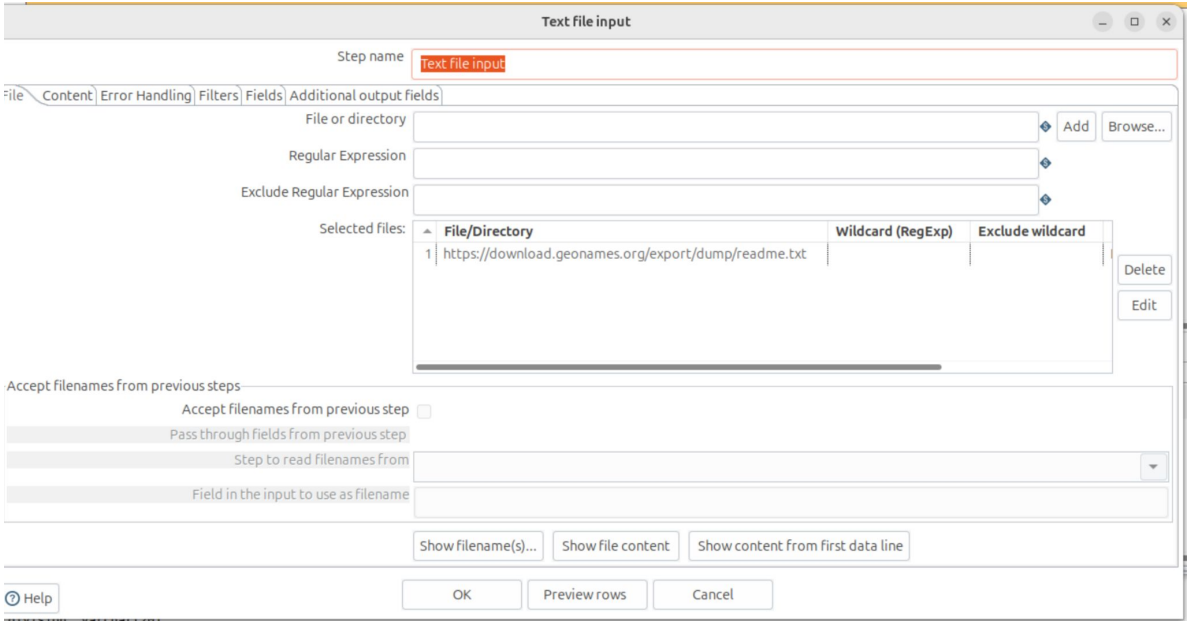
Obtener datos a través de la web mediante:

- *Text-Based Web Extraction*
- *HTTP Client*
- *Web services*
  - *REST client*
  - *Web services lookup - SOAP*



Extracción basada en web

Desde el paso ***Text input data*** se puede utilizar una URL como nombre de archivo que utiliza Apache VFS.



Extracción basada en web

HTTP Client

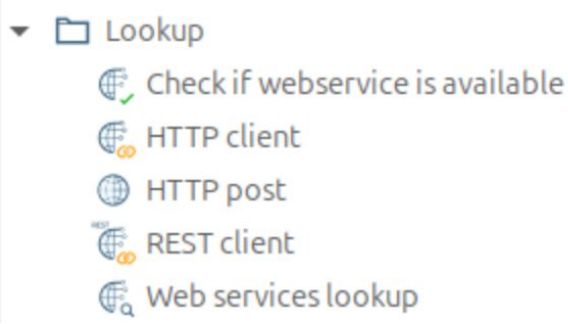
Recuperación de datos estructurados de Internet.

Realiza una llamada a la URL específica y retorna los datos como String.

Este resultado puede ser texto normal que se puede guardar como un archivo *delimited*, por ejemplo CSV o en XML.

El resultado debe ser procesado posteriormente dependiendo del formato de salida.

Es un paso de búsqueda y necesita una entrada para activarse.



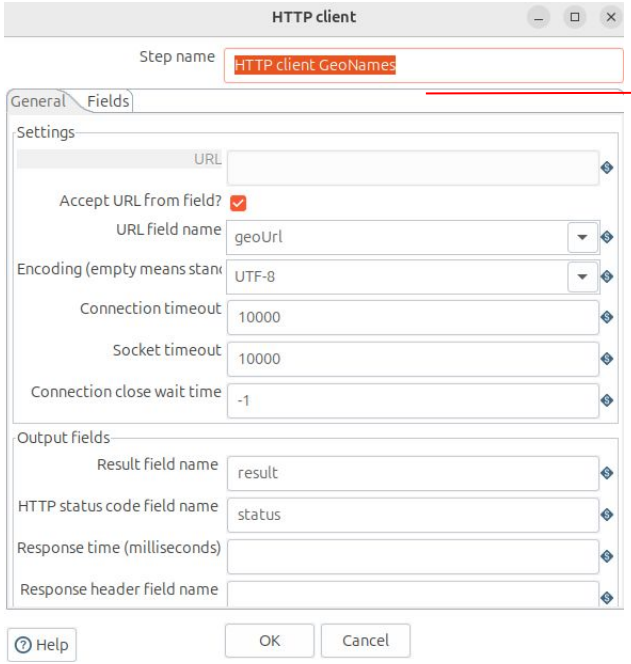
Extracción basada en web

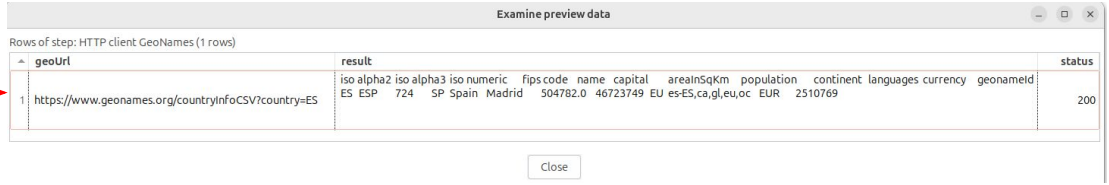
HTTP Client

f1	f2	f3	f4	f5	f6	f7	f8	f9	f10	f11	f12
iso alpha2	iso alpha3	iso numeric	fips code	name	capital	areaInSqKm	population	continent	languages	currency	geonameId
ES	ESP	724	SP	Spain	Madrid	504782.0	46723749	EU	es-ES,ca,gl,eu,oc	EUR	2510769

Extracción basada en web

HTTP Client



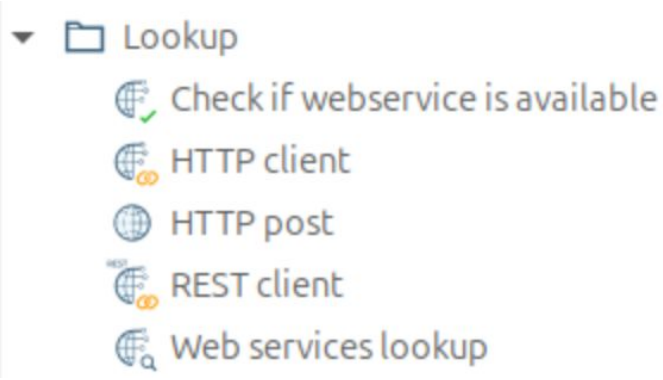


HTTP client GeoNames

https://www.geonames.org/countryInfoCSV?country=ES

Extracción basada en web

Web services



REST client

- REST (Representational State Transfer)

Web services lookup

- SOAP (Simple Object Access Protocol)
- Utiliza the Web Services Description Language (WSDL)

Extracción basada en web

Web services REST client

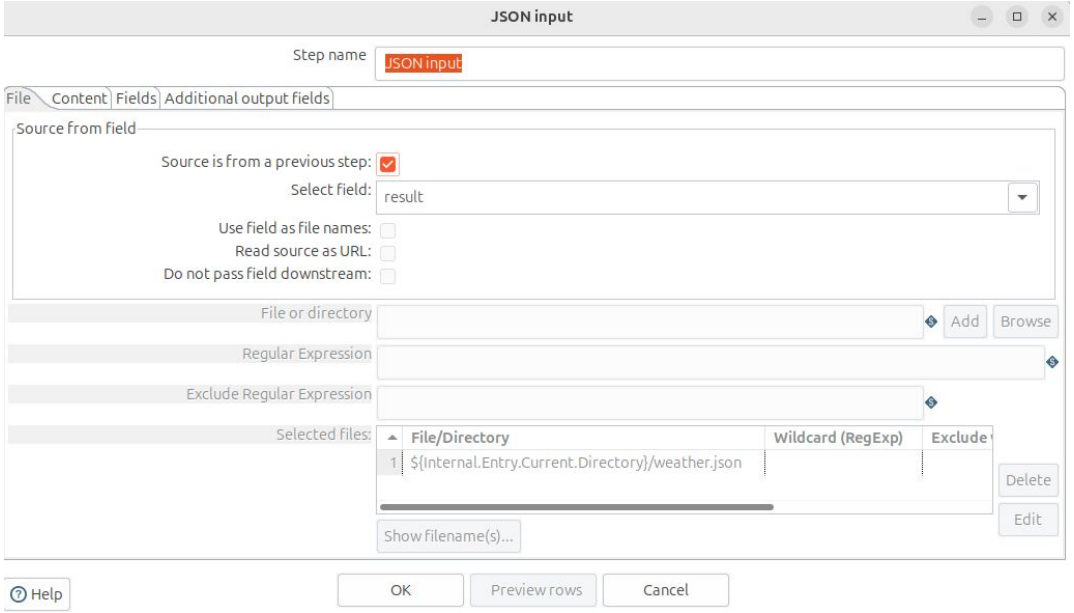
OpenWeather

https://api.openweathermap.org/data/2.5/weather?q=Zaragoza&APPID=


```
{
  "coord": {
    "lon": -0.8773,
    "lat": 41.6561
  },
  "weather": [
    {
      "id": 800,
      "main": "Clear",
      "description": "clear sky",
      "icon": "01d"
    }
  ],
  "base": "stations",
  "main": {
    "temp": 302.75,
    "feels_like": 301,
    "temp_min": 302.4,
    "temp_max": 303.16,
    "pressure": 1014,
    "humidity": 17,
    "sea_level": 1014,
    "grnd_level": 976
  },
  "visibility": 10000,
  "wind": {
    "speed": 6.69,
    "deg": 310
  },
  "clouds": {
    "all": 0
  },
  "dt": 1757006653,
  "sys": {
    "type": 2,
    "id": 2003310,
    "country": "ES",
    "sunrise": 1756963963,
    "sunset": 1757010766
  },
  "timezone": 7200,
  "id": 3104324,
  "name": "Zaragoza",
  "cod": 200
}
```

Extracción basada en web

Web services REST client



Extracción basada en web

Web services lookup

Para servicios web basados en Simple Object Access Protocol (SOAP)

Utiliza Web Services Description Language (WSDL) para cargar las operaciones y mapear campos.

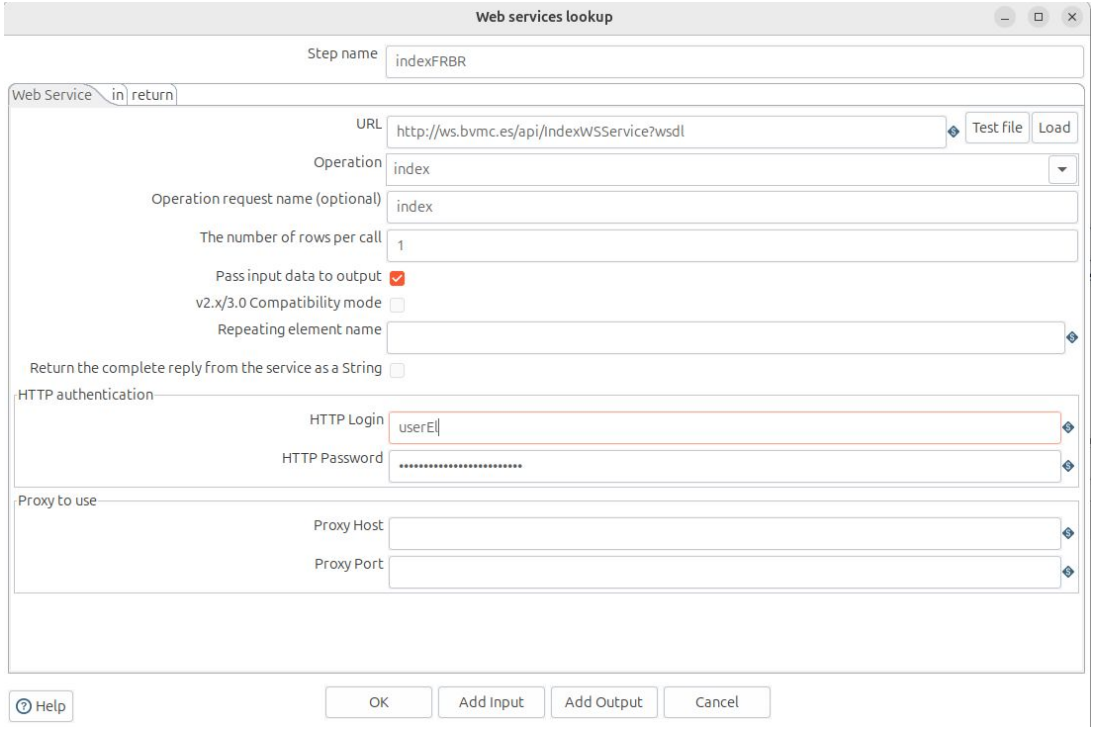
En URL se indica la dirección del WSDL

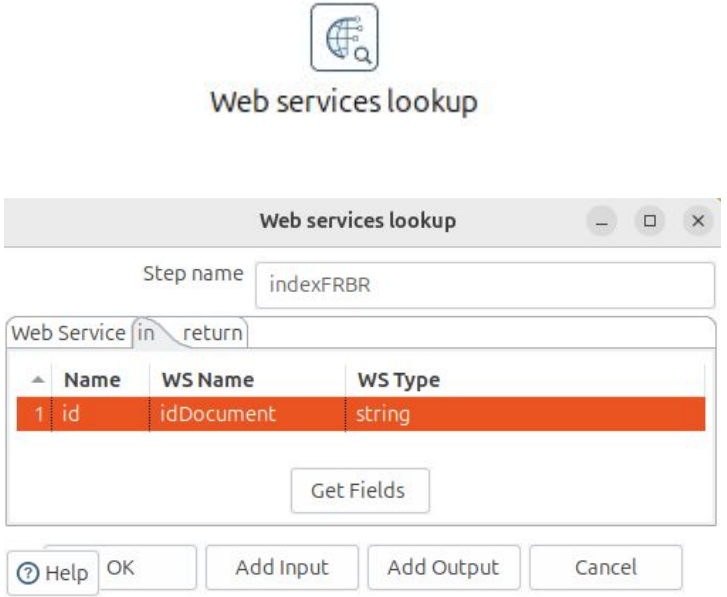
Puede necesitar usuario y contraseña

No necesita ser activado por un paso previo como en el caso de HTTP Client.

Extracción basada en web

Web services lookup





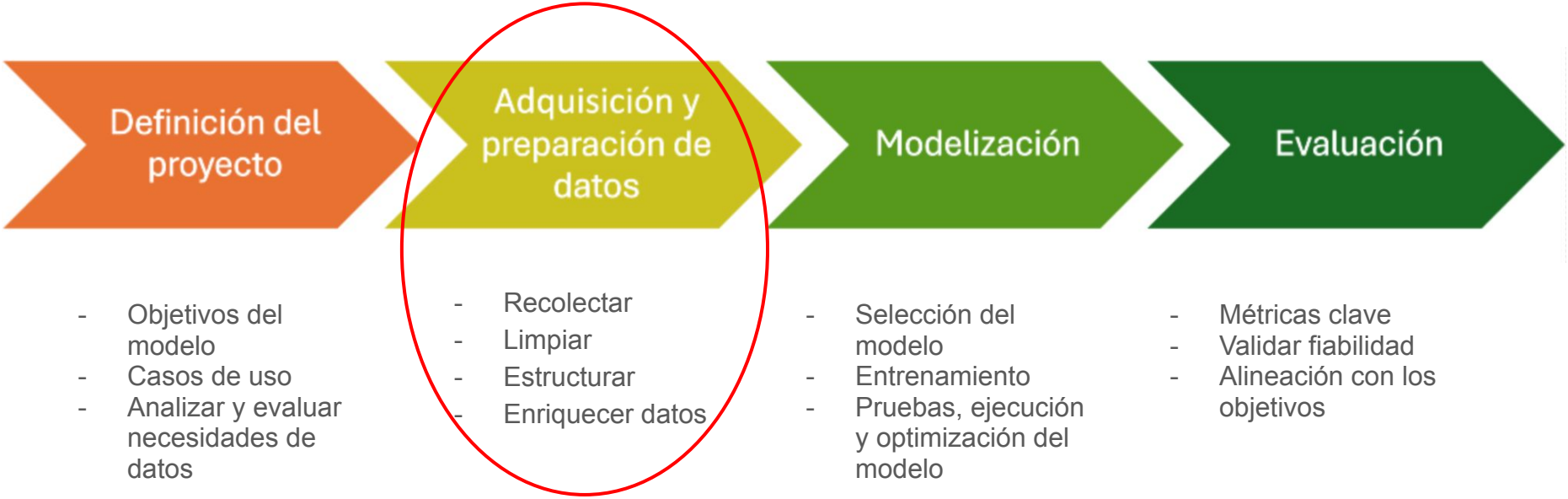
Procesos ETL

Tema 6: Limpieza, transformación y normalización de datos. Parte I Expresiones regulares

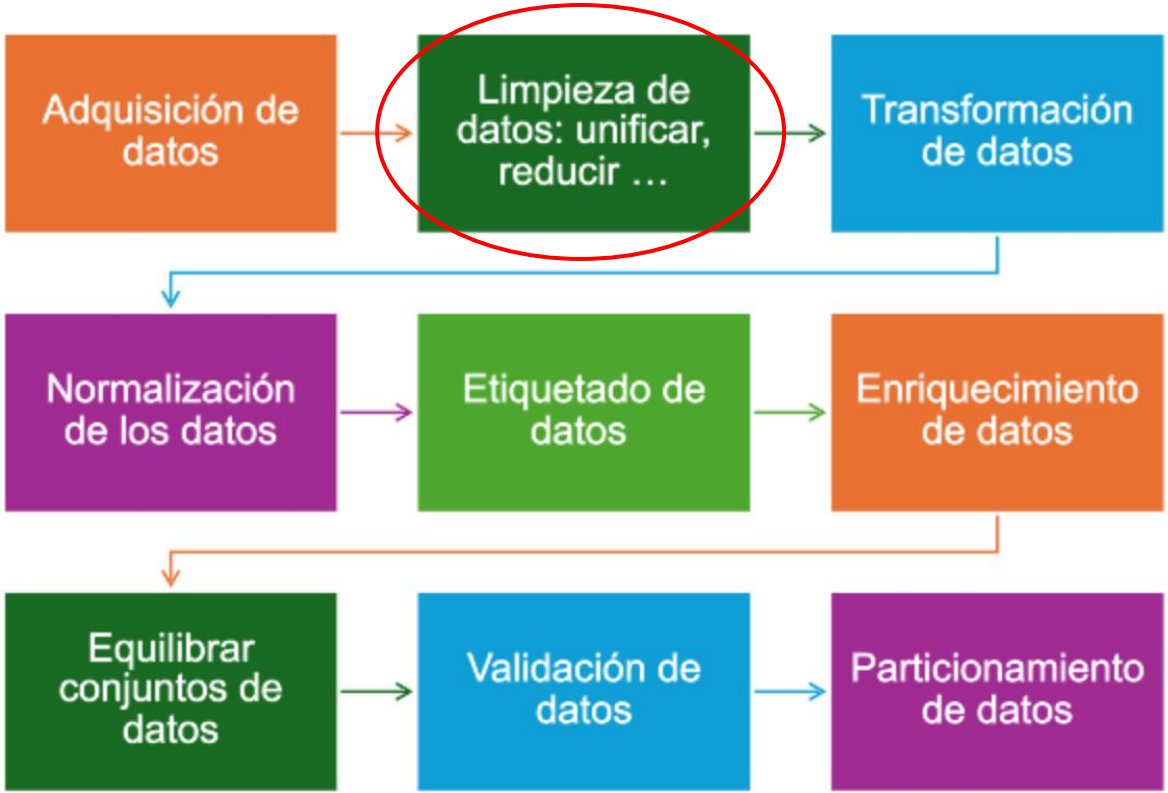
Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Ciclo de vida del proyecto IA



Preparación de datos



Qué es un proceso ETL

Extract

- **Extraer** datos desde las distintas fuentes de origen - fusión de datos
  - Identificar fuentes
  - Identificar conjunto de datos
  - Conectores para la adquisición
  - Fuentes como Hadoop, datos abiertos, Sql, Nosql (MongoDB, Cassandra), repositorios RDF, Wikidata, etc
- **Analizar** los datos obtenidos
  - Tipos de datos
  - Codificación origen
  - Semántica
- **Interpretar** y comprobar que cumplen con la estructura esperada

Qué es un proceso ETL

Transform

Depurar y ajustar a los requisitos, aplicando **reglas** que guíen las transformaciones de los datos para un formato final, contenido y forma

- Selección de datos (columnas, filas…), eliminado ruido
- Modificación del contenido o la estructura de los datos
- Integración con datos de otras fuentes
- Corrección de errores
- Cálculo de valores derivados o agregados a partir de datos procesados
- Traducción de valores
- Codificación
- Tratamiento de textos
- Validación de datos frente a reglas de calidad

Qué es un proceso ETL

Load

Carga y actualización de los datos en el almacén de datos

- Selección sistema destino: fichero, *data warehouse*, BD relacional, índice, NoSql, etc
- Nuevo o sobrescribir (ejemplo BD, Excel)
- Historial de cambios

Tipos de proceso de carga:

- **Acumulación simple**
- **Rolling** para mayor granularidad

Limpieza de datos

Limpieza de datos

Los datos sin procesar suelen ser imperfectos, contienen errores, faltan valores o están duplicados.

Objetivo de la limpieza de datos es **eliminar** estas **anomalías para garantizar la calidad y su fiabilidad**.

Una limpieza adecuada ayuda a evitar que los datos defectuosos afecten al rendimiento del modelo.

- Corregir **errores**
- Eliminar **duplicados**
- Gestionar valores que **faltan**: reemplazo, interpolación o eliminación
- Gestionar valores **atípicos**
- **Unificar** de tipos de datos
- Aplicar la **reducción** de datos
    - La reducción de datos filtra los campos que no son necesarios
    - Eliminar los valores que se encuentran fuera del rango esperado
    - Redundancia, conservar datos redundantes puede ralentizar el análisis y consumir recursos

Limpieza de datos

Limpieza de datos mediante:

1. Expresiones regulares
   - Introducción
   - Sintaxis
   - Ejemplos

2. Otros mecanismos para la limpieza de datos

Expresiones regulares

Expresiones regulares

Mecanismo **eficiente y flexible** que facilita el procesamiento de datos.

Secuencia de caracteres que forman un patrón que facilitan **filtrar, encontrar, validar y tratar** un texto.

Sus principales componentes:

- **Literales**: caracteres que conforman cualquier palabra
- **Clases** de caracteres: son definidas a través de una lista de caracteres entre corchetes
- **Metacaracteres**: elemento especial que permiten delimitar, iterar, alternar a los literales.

Expresiones regulares

Regular Expressions Cheat Sheet by Dave Child (DaveChild) via cheatography.com/1/cs/5/

https://cheatography.com/davechild/cheat-sheets/regular-expressions/pdf/

Cheatography

Regular Expressions Cheat Sheet
by Dave Child (DaveChild) via cheatography.com/1/cs/5/

Anchors

^	Start of string, or start of line in multi-line pattern
\A	Start of string
$	End of string, or end of line in multi-line pattern
\Z	End of string
\b	Word boundary
\B	Not word boundary
\<	Start of word
\>	End of word

Character Classes

\c	Control character
\s	White space
\S	Not white space
\d	Digit
\D	Not digit
\w	Word
\W	Not word
\x	Hexadecimal digit
\O	Octal digit

POSIX

[:upper:]	Upper case letters
[:lower:]	Lower case letters
[:alpha:]	All letters
[:alnum:]	Digits and letters
[:digit:]	Digits
[:xdigit:]	Hexadecimal digits
[:punct:]	Punctuation
[:blank:]	Space and tab
[:space:]	Blank characters
[:cntrl:]	Control characters
[:graph:]	Printed characters
[:print:]	Printed characters and spaces
[:word:]	Digits, letters and underscore

Assertions

?=	Lookahead assertion
?!	Negative lookahead
?<=	Lookbehind assertion
?!= or ?<!	Negative lookbehind
?>	Once-only Subexpression
?()	Condition [if then]
?()|	Condition [if then else]
?#	Comment

Quantifiers

*	0 or more	{3}	Exactly 3
+	1 or more	{3,}	3 or more
?	0 or 1	{3,5}	3, 4 or 5

Add a ? to a quantifier to make it ungreedy.

Escape Sequences

\	Escape following character
\Q	Begin literal sequence
\E	End literal sequence

"Escaping" is a way of treating characters which have a special meaning in regular expressions literally, rather than as special characters.

Common Metacharacters

^	[	.	$
{	*	(	\
+	)	|	?
<	>		

The escape character is usually \

Special Characters

\n	New line
\r	Carriage return
\t	Tab
\v	Vertical tab
\f	Form feed
\xxx	Octal character xxx
\xhh	Hex character hh

Groups and Ranges

.	Any character except new line (\n)
(a|b)	a or b
(...)	Group
(?:...)	Passive (non-capturing) group
[abc]	Range (a or b or c)
[^abc]	Not (a or b or c)
[a-q]	Lower case letter from a to q
[A-Q]	Upper case letter from A to Q
[0-7]	Digit from 0 to 7
\x	Group/subpattern number "x"

Ranges are inclusive.

Pattern Modifiers

g	Global match
i *	Case-insensitive
m *	Multiple lines
s *	Treat string as single line
x *	Allow comments and whitespace in pattern
e *	Evaluate replacement
U *	Ungreedy pattern

* PCRE modifier

String Replacement

$n	nth non-passive group
$2	"xyz" in /^(abc(xyz))$/
$1	"xyz" in /^(?:abc)(xyz)$/
$`	Before matched string
$'	After matched string
$+	Last matched string
$&	Entire matched string

Some regex implementations use \ instead of $.

By **Dave Child** (DaveChild)
cheatography.com/davechild/
www.addedbytes.com

Published 19th October, 2011.
Last updated 2nd January, 2015.
Page 1 of 1.

Sponsored by **Readability-Score.com**
Measure your website readability!
https://readability-score.com

Expresiones regulares

Consultar RegExLib.com Regular Expression Cheat Sheet

regexlib.com/CheatSheet.aspx

RegExLib.com Regular Expression Library

RegExLib.com Regular Expression Cheat Sheet (.NET)

Metacharacters Defined		Metacharacter Examples	
^	Start of a string.	^abc	abc, abcdefg, abc123, ...
$	End of a string.	abc$	abc, endsinabc, 123abc, ...
.	Any character (except \n newline)	a.c	abc, aac, acc, adc, aec, ...
|	Alternation.	bill|ted	ted, bill
{...}	Explicit quantifier notation.	ab{2}c	abbc
[...]	Explicit set of characters to match.	a[bB]c	abc, aBc
(...)	Logical grouping of part of an expression.	(abc){2}	abcabc
*	0 or more of previous expression.	ab*c	ac, abc, abbc, abbbc, ...
+	1 or more of previous expression.	ab+c	abc, abbc, abbbc, ...
?	0 or 1 of previous expression; also forces minimal matching when an expression might match several strings within a search string.	ab?c	ac, abc
\	Preceding one of the above, it makes it a literal instead of a special character. Preceding a special matching character, see below.	a\sc	a c

Character Escapes http://tinyurl.com/5wm3wl

ordinary characters	Characters other than . $ ^ { [(|)] } * + ? \ match themselves.
\a	Matches a bell (alarm) \u0007.
\b	Matches a backspace \u0008 if in a []; otherwise matches a word boundary (between \w and \W characters).
\t	Matches a tab \u0009.
\r	Matches a carriage return \u000D.
\v	Matches a vertical tab \u000B.
\f	Matches a form feed \u000C.
\n	Matches a new line \u000A.
\e	Matches an escape \u001B.
\040	Matches an ASCII character as octal (up to three digits); numbers with no leading zero are backreferences if they have only one digit or if they correspond to a capturing group number. (For more information, see Backreferences.) For example, the character \040 represents a space.
\x20	Matches an ASCII character using hexadecimal representation (exactly two digits).
\cC	Matches an ASCII control character; for example \cC is control-C.
\u0020	Matches a Unicode character using a hexadecimal representation (exactly four digits).
*	When followed by a character that is not recognized as an escaped character, matches that character. For example, ***** is the same as **\x2A**.

Character Classes http://tinyurl.com/5ck4ll

.	Matches any character except \n. If modified by the Singleline option, a period character matches any character. For more information, see Regular Expression Options.
[aeiou]	Matches any single character included in the specified set of characters.
[^aeiou]	Matches any single character not in the specified set of characters.
[0-9a-fA-F]	Use of a hyphen (–) allows specification of contiguous character ranges.
\p{name}	Matches any character in the named character class specified by {name}. Supported names are Unicode groups and block ranges. For example, Ll, Nd, Z, IsGreek, IsBoxDrawing.
\P{name}	Matches text not included in groups and block ranges specified in {name}.
\w	Matches any word character. Equivalent to the Unicode character categories [\p{Ll}\p{Lu}\p{Lt}\p{Lo}\p{Nd}\p{Pc}]. If ECMAScript-compliant behavior is specified with the ECMAScript option, \w is equivalent to [a-zA-Z_0-9].
\W	Matches any nonword character. Equivalent to the Unicode categories [^\p{Ll}\p{Lu}\p{Lt}\p{Lo}\p{Nd}\p{Pc}]. If ECMAScript-compliant behavior is specified with the ECMAScript option, \W is equivalent to [^a-zA-Z_0-9].
\s	Matches any white-space character. Equivalent to the Unicode character categories [\f\n\r\t\v\x85\p{Z}]. If ECMAScript-compliant behavior is specified with the ECMAScript option, \s is equivalent to [\f\n\r\t\v].
\S	Matches any non-white-space character. Equivalent to the Unicode character categories [^\f\n\r\t\v\x85\p{Z}]. If ECMAScript-compliant behavior is specified with the ECMAScript option, \S is equivalent to [^ \f\n\r\t\v].
\d	Matches any decimal digit. Equivalent to \p{Nd} for Unicode and [0-9] for non-Unicode, ECMAScript behavior.

Expresiones regulares

Algunos de los más utilizados *metacharacters*:

- **** Indica que el siguiente carácter es especial y se desea tratar como un literal
- **^** Indica el principio de una cadena
- **$** Indica el final de una cadena
- **.** Cualquier carácter salvo el salto de línea (\n)
- **(...)** Agrupación lógica de partes de una expresión.
- **[...]** Un conjunto de caracteres de la expresión
  - [xyz] Indica coincidencia con cualquiera de los caracteres entre corchetes
  - [^xyz] cualquiera carácter que no esté entre corchetes
  - [a-z] cualquier carácter dentro del rango especificado
  - [0-9a-fA-F] cualquier carácter dentro de los rangos especificados
- **{...}** Indica un número o intervalo de longitud de la expresión
  - {n} repetición de un carácter n veces.
  - {n,} repetición de un carácter como mínimo n veces
  - {n,m} repetición de un carácter como mínimo n veces y máximo m veces
- **?** 0-1 ocurrencias de la expresión
- **+** 1-n ocurrencias de la expresión
- ***** 0-n ocurrencias de la expresión
- | Para indicar una disyunción lógica, elegir entre dos valores x|y se tiene que cumplir al menos uno de los dos

Expresiones regulares

Algunos de los más utilizados *character classes*:

- **\w** Coincide con cualquier carácter alfanumérico o _, equivale a [a-zA-Z_0-9]
- **\W** Coincide con cualquier carácter no alfanumérico ni _, equivale a [^a-zA-Z_0-9]
- **\s** Espacio en blanco, ECMAScript equivale a [\f\n\r\t\v]
- **\S** No espacio en blanco, ECMAScript equivale a [^ \f\n\r\t\v]
- **\d** Dígito decimal, equivale a [0-9]
- **\D** No dígito decimal, equivale a [^0-9]
- **\x** Dígito hexadecimal
- **\p**{name} Coincide con cualquier carácter de la clase o categoría Unicode especificada por {name}.
- **\P**{name} Coincide con el texto no incluido en los grupos y rangos de bloques especificados en {name}

Anchors:

- **\b** Límite de palabra
- **\B** No límite de palabra

Expresiones regulares

Flags, modifican el comportamiento del patrón

- **g** encuentra todas las coincidencias, si no se usa solo devuelve la primera coincidencia
- **i** no distingue entre mayúsculas y minúsculas
- **m** Para múltiples líneas
- **s** Habilita el modo “dotall”, un punto “.” coincide con el carácter de línea nueva \n tratando a la cadena como una única línea
- **y** búsqueda en la posición exacta del texto
- **u** Permite el soporte completo de Unicode

Expresiones regulares

PDI incluye pasos para el uso de expresiones regulares.

Facilita la inclusión de scripts en Javascript para la limpieza de datos y otras tareas.

En JavaScript se pueden utilizar los métodos `exec()` y `test()` de `RegExp`, y `match()`, `matchAll()`, `replace()`, `replaceAll()`, `search()` y `split()` de `String`.

```
str = "Impreso en casa de Monreal núm. 2, 1902";
const regex = /\d+$/;
console.log(regex.test(str)); // true

console.log(regex.exec(str));
```

```
[
  '1902',
  index: 35,
  input: 'Impreso en casa de Monreal núm. 2, 1902',
  groups: undefined
]
```

```
const regex2 = /\d+/g;
const result = Array.from(str.matchAll(regex2), (m) => m[0]);

console.log(result);
```

```
[ '2', '1902' ]
```

https://developer.mozilla.org/en-US/docs/Web/JavaScript/Reference/Regular_expressions

Expresiones regulares

^ $ Inicio o final de un string (*anchors*)

- ^(.+:) o bien ^(Auxiliar localización):?( )?

```
s = "Auxiliar localización: Parque Natural de Redes, Asturias";
const pRE = /^(Auxiliar localización):?( )?.*/;
pRE.test(s);
```

- (.*)(\d+)$ cadena que termine en número

```
s = "Auxiliar localización: Parque Natural de Redes, Asturias, 1968"
const pRE = /.*\d+$/;
pRE.test(s);
```

- Ambos se puede usar juntos para comprobar si una cadena coincide completamente con el patrón

```
s = "09:37";
const pRE = /^\d\d:\d\d$/;
pRE.test(s);
```

Expresiones regulares

- **\b** Búsqueda de palabras completas

```
s1 = "Edición digital a partir de Obras completas de Baltasar Gracián. Vol II";
s2 = "Ed. Emilio Blanco sin Volumen definido. 1993";

const pRE = /\bVol\b/;
pRE.test(s1);
pRE.test(s2);

console.log(s1.replace(pRE, 'Volumen'));
console.log(s2.replace(pRE, 'Volumen'));

const pRE2 = /Vol/;
pRE2.test(s1);
pRE2.test(s2);

console.log(s1.replace(pRE2, 'Volumen'));
console.log(s2.replace(pRE2, 'Volumen'));
```

Expresiones regulares

- \ Indica que el siguiente carácter es especial y se desea tratar como un literal
  - Para escapar caracteres especiales [\ ^ $. | ? * + ( )
  - Se debe escapar / si se usa el patrón /.../ (pero no dentro de new RegExp).
  - Si se usa new RegExp, se deben duplicar las barras invertidas \\, porque las comillas de cadena consumen una.

```
s1 = "Edición digital a partir de Obras completas de Baltasar Gracián. Volumen II";
s2 = "Ed. Emilio Blanco sin Volumen, (Murcia)";

const pRE = /\./;
pRE.test(s1);

const pRE2 = /(.*, )?\(?(.+)\)?$/;
pRE2.test(s1);
pRE2.test(s2);
```

Expresiones regulares

- **()** Para agrupar

```
s1 = "Auxiliar localización: Sanlúcar de Barrameda, Cádiz";
s2 = "Ed. Emilio Blanco sin Volumen, (Murcia)";

const pRE = /(.*, )?\(?(.+)\)?$/;
pRE.test(s1); // → true grupo2 = Cádiz
pRE.test(s2); // → true grupo2 = Murcia
```

Expresiones regulares

- **[...]** Un conjunto de caracteres de la expresión
  - [xyz] Indica coincidencia con cualquiera de los caracteres entre corchetes
  - [^xyz] cualquiera caracter que no esté entre corchetes
  - [a-z] cualquier caracter dentro del rango especificado
  - [0-9a-fA-F] cualquier caracter dentro de los rangos especificados

```
//Obtener todas las obras cuya reproducción indique un volumen. Existen diferentes formas de indicarlo

s1 = "Edición digital a partir de Obras completas de Baltasar Gracián. Volumen II";
s2 = "Ed. Emilio Blanco volumen I, pp23, (Murcia)";
s3 = "Ed. Emilio Blanco Vol.2 (Murcia)";


const pRE = /[Vv]olumen/;
pRE.test(s1);
pRE.test(s2);
pRE.test(s3);
```

```
const pRE2 = /[Vv]ol(\.|umen)/;
pRE2.test(s1));
pRE2.test(s2));
pRE2.test(s3));
```

Expresiones regulares

```
//Obtener todas las obras cuya reproducción indique un volumen. Existen diferentes formas de indicarlo

s1 = "Edición digital a partir de Obras completas de Baltasar Gracián. Volumen II";
s2 = "Ed. Emilio Blanco volumen I, pp23, (Murcia)";
s3 = "Ed. Emilio Blanco Vol.2 (Murcia)";


 const pRE = /[Vv]olumen/;                                const pRE2 = /[Vv]ol(\.|umen)/;
 pRE.test(s1);                                            console.log(pRE2.test(s1));
 pRE.test(s2);                                            console.log(pRE2.test(s2));
 pRE.test(s3);                                            console.log(pRE2.test(s3));
```

¿Alguna otra expresión regular que podamos usar?

Expresiones regulares

```
//Obtener todas las obras cuya reproducción indique un volumen. Existen diferentes formas de indicarlo

s1 = "Edición digital a partir de Obras completas de Baltasar Gracián. Volumen II";
s2 = "Ed. Emilio Blanco volumen I, pp23, (Murcia)";
s3 = "Ed. Emilio Blanco Vol.2 (Murcia)";


 const pRE = /[Vv]olumen/;
 pRE.test(s1); // → true
 pRE.test(s2); // → true
 pRE.test(s3); // → false

const pRE2 = /[Vv]ol(\.|umen)/;
pRE2.test(s1); // → true
pRE2.test(s2); // → true
pRE2.test(s3); // → true
```

/\bVol(\.|umen)?\b/i

Expresiones regulares

- **[...]** Un conjunto de caracteres de la expresión

Las clases de caracteres son abreviaturas para ciertos conjuntos de caracteres:

\d →[0-9]
\w →[a-zA-Z0-9_]
\s →[\t\n\v\f\r], además de otros caracteres de espacio de unicode.

Expresiones regulares

Ejercicio 1 - Obtener expresión regular para comprobar que una fecha del siglo 21 es correcta en formato YYYY-MM-DD

```
s1 = "[Vicente Barberá Masip o Enrique Desfilis Barberá]., 1898";
s2 = "Imp. Vda. de Ayoldi, 1883-06-24";
s3 = "Direcció General del Llibre, Arxius i Biblioteques, 2022-12-31";


const pRE = /…/;


console.log(pRE.test(s1));
console.log(pRE.test(s2));
console.log(pRE.test(s3));
```

Expresiones regulares

Ejercicio 1 - Obtener expresión regular para comprobar que una fecha del siglo 21 es correcta en formato YYYY-MM-DD

```
s1 = "[Vicente Barberá Masip o Enrique Desfilis Barberá]., 1898";
s2 = "Imp. Vda. de Ayoldi, 1883-06-24";
s3 = "Direcció General del Llibre, Arxius i Biblioteques, 2022-12-31";


const pRE = /(20\d{2}|2100)-(0[1-9]|1[0-2])-(0[1-9]|[12]\d|3[01])/;
console.log(pRE.test(s1)); // → false
console.log(pRE.test(s2)); // → false
console.log(pRE.test(s3)); // → true


// (20(1\d|2[0-5]) -> Años del 2000 al 2025.
// (0[1-9]|1[0-2]) -> Grupo de captura el mes (01-12)
// (0[1-9]|[12]\d|3[01]) -> Grupo de captura para el día (01-31)
```

Expresiones regulares

- . Cualquier carácter salvo el salto de línea
- + 1-n ocurrencias de la expresión
- * 0-n ocurrencias de la expresión
- ? 0-1 ocurrencias de la expresión

```
s1 = "Auxiliar localización: Sanlúcar de Barrameda, Cádiz";
s2 = "Ed. Emilio Blanco sin Volumen, (Murcia)";
s3="";

pRE = /.*/;
console.log(pRE.test(s1)); // → true
console.log(pRE.test(s3)); // → ?

pRE = /.+/;
console.log(pRE.test(s3)); // → ?

pRE = /(.*, )?\(?(.+)\)?$/;
pRE.test(s1); // → ?
pRE.test(s2); // → ?
```

Expresiones regulares

- **{...}** Indica un número o intervalo de longitud de la expresión
  - {n} repetición de un carácter n veces.
  - {n,} repetición de un carácter como mínimo n veces
  - {n,m} repetición de un carácter como mínimo n veces y máximo m veces

```
codigos = "CP: 28001, 03001, 123456.";
console.log(codigos.match(/\b\d{5}\b/g));

//Extrae número de registro variable
idReg = "ID123, ID4567, ID7450123.";
validados = idReg.match(/\d{4,6}/g);
```

Expresiones regulares

Ejercicio 2 - Validar una contraseña mediante una expresión regular. La contraseña debe tener entre 8 y 16 caracteres, contener al menos una letra minúscula, una mayúscula y un número.

Expresiones regulares

Ejercicio 2 - Validar una contraseña mediante una expresión regular. La contraseña debe tener entre 8 y 16 caracteres y contener al menos una letra minúscula, una mayúscula y un número.

```
/**
 * Validar contraseña con una expresión regular.
 * @param {string}
 * @returns {boolean} true si es válida y false si no lo es
 */
function isValid(pass) {
  const regex = /^(?=.*[a-z])(?=.*[A-Z])(?=.*\d).{8,16}$/;
  return regex.test(pass);
}


console.log(isValid("P@ss11111!"));
console.log(isValid("NoSeQue_poner3"));
console.log(isValid("1234567EsoEs"));
console.log(isValid("1234"));
```

Expresiones regulares

Ejercicio 3 - ¿Cómo limpiamos dos o más espacios en blanco en la siguiente cadena de texto?

Es el título original de una obra publicada en la BVMC, tal cual se puede leer a continuación:

Recopilacion verificada por Felipe Mateu, de las noticias y reglas por las que se gobierna la Real Acequia de Moncada (Valencia), efectuada por acuerdo de su Junta de cequeiros en 10 de octubre de 1757 : Concluida esta recopilación en 16 de julio de 1758

Expresiones regulares

Ejercicio 3:

```
s1 = "Recopilacion verificada por Felipe Mateu, de las noticias y reglas por las que se gobierna la
Real Acequia de Moncada (Valencia), efectuada por acuerdo de su Junta de cequeiros en 10 de octubre de
1757  : Concluida esta recopilación en 16 de julio de 1758";

const pRE = /( ){2,}/;
const pRE2 = /\s{2,}/;
const pRE3 = /\s+/g;
const pRE4 = /\s{2,}/g;

console.log(pRE.test(s1));
console.log(pRE2.test(s1));
console.log(pRE3.test(s1));
console.log(pRE4.test(s1));

let s1SinEspaciosMultiples = s1.replace(/\s+/g, ' ');

console.log(pRE4.test(s1SinEspaciosMultiples));
```

Expresiones regulares

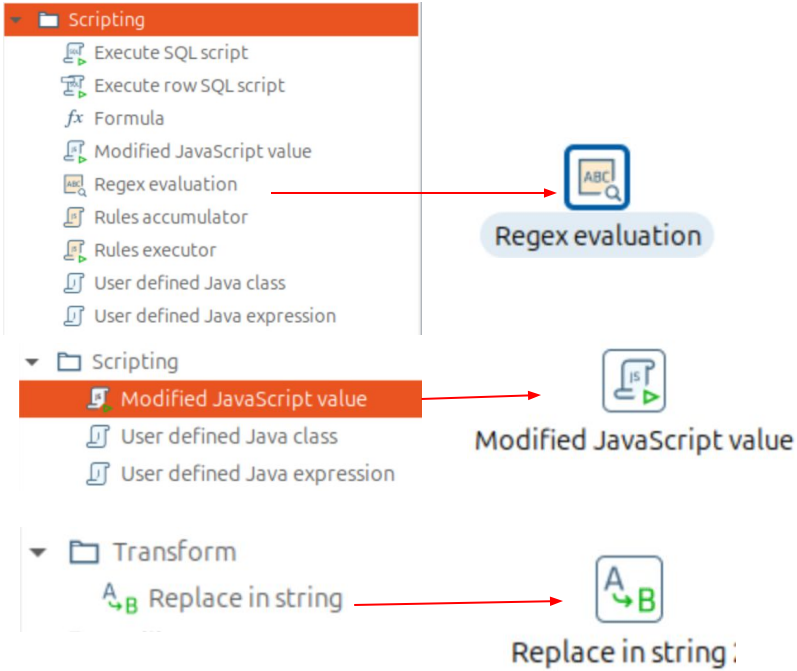
Ejercicio 3:

```
s1 = "Recopilacion verificada por Felipe Mateu, de las noticias y reglas por las que se gobierna la
Real Acequia de Moncada (Valencia), efectuada por acuerdo de su Junta de cequeiros en 10 de octubre de
1757  : Concluida esta recopilación en 16 de julio de 1758";

const pRE = /( ){2,}/; //encuentra al menos dos espacios en blanco seguidos
const pRE2 = /\s{2,}/; //encuentra al menos dos espacios en blanco seguidos
const pRE3 = /\s+/g; //encuentra espacios en blanco lo aplica en global
const pRE4 = /\s{2,}/g;  //encuentra al menos dos espacios en blanco seguidos lo aplica en global

console.log(pRE.test(s1)); // → true
console.log(pRE2.test(s1)); // → true
console.log(pRE3.test(s1)); // → true
console.log(pRE4.test(s1)); // → true

// método replace() con la expresión regular
let s1SinEspaciosMultiples = s1.replace(/\s+/g, ' ');
console.log(s1SinEspaciosMultiples);

console.log(pRE4.test(s1SinEspaciosMultiples)); //false
```

Expresiones regulares en PDI

Ejemplos de uso en Pentaho Data Integration

- Reemplazar partes en una cadena de texto ***replace string***
- Obtener datos de un texto mediante **regex evaluation**
- Expresiones regulares **REGEX** en MySql
- **JavaScript**



Expresiones regulares en PDI

Replace string

Se debe limpiar los campos de unos registros bibliográficos obtenidos de una fuente externa.

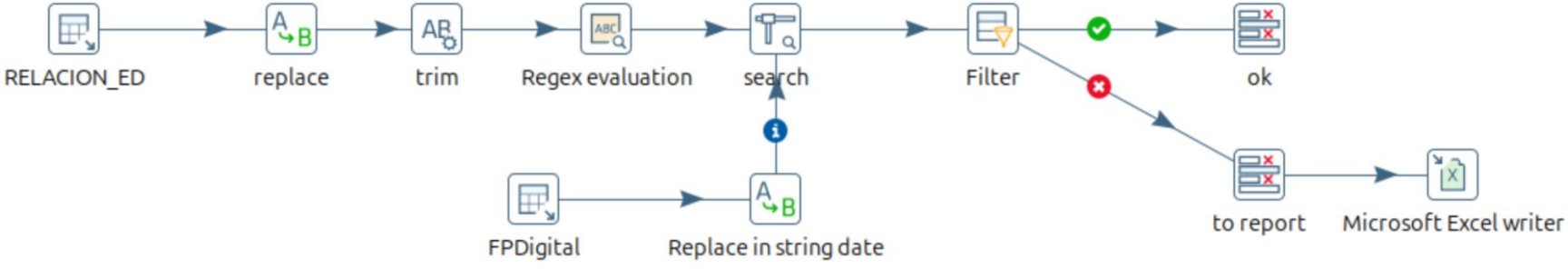
Después de analizar una muestra, se determina que es necesario adecuar a la norma de catalogación y forma de descripción correcta para el almacén de datos final.

Para ello se solicita realizar un proceso ETL para dicha tarea.

mencionPublicacion
"Cuadernos Hispanoamericanos", núm. 91-92 (julio-agosto 1957), pp. 180-201
"Cuadernos Hispanoamericanos", núm. 45 (septiembre 1953), pp. 279-291
"Cuadernos Hispanoamericanos", núm. 22 (julio-agosto 1951), pp. 15-19
"Cuadernos Hispanoamericanos", núm. 26 (febrero 1952)
"Cuadernos Hispanoamericanos", núm. 20 (marzo-abril 1951), pp. 169-185
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 237-238
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 241-242
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 243-286
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 289-299
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 301-306
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 323-332
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 335-381
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 383-397
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 399-414
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 415-417
"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 421-426

Expresiones regulares en PDI

Replace string



757355	"Cuadernos Hispanoamericanos", núm. 91-92 julio-agosto 1957	"Cuadernos Hispanoamericanos", núm. 91-92 (julio-agosto 1957), pp. 180-201
782746	"Cuadernos Hispanoamericanos", núm. 45 septiembre 1953	"Cuadernos Hispanoamericanos", núm. 45 (septiembre 1953), pp. 279-291
970729	"Cuadernos Hispanoamericanos", núm. 22 julio-agosto 1951	"Cuadernos Hispanoamericanos", núm. 22 (julio-agosto 1951), pp. 15-19
972802	"Cuadernos Hispanoamericanos", núm. 26 febrero 1952	"Cuadernos Hispanoamericanos", núm. 26 (febrero 1952)
975986	"Cuadernos Hispanoamericanos", núm. 20 marzo-abril 1951	"Cuadernos Hispanoamericanos", núm. 20 (marzo-abril 1951), pp. 169-185
984341	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 237-238
984389	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 241-242
984392	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 243-286
984397	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 289-299
984401	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 301-306
984407	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 323-332
984413	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 335-381
984449	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 383-397
984452	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 399-414
984455	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 415-417
984461	"Cuadernos Hispanoamericanos", núm. 11-12 septiembre-diciembre 1949	"Cuadernos Hispanoamericanos", núm. 11-12 (septiembre-diciembre 1949), pp. 421-426

Expresiones regulares en PDI

Replace string

Replace in string

Step name: replace

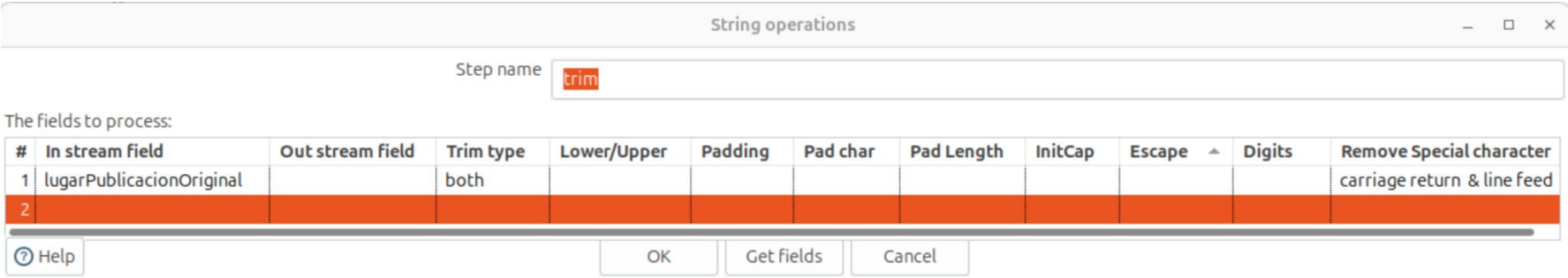
Fields string

	In stream field	O	use RegEx	Search	Replace with	Set empty string?	Replace with field	Whole Word	Case sensitive	Is Unicode
1	lugarPublicacionOriginal		Y	, pp\. .*$		N		N	N	N
2	lugarPublicacionOriginal		Y	, p\. .*$		N		N	N	N
3	lugarPublicacionOriginal		Y	(\)|\]|\?)$		N		N	N	N
4	lugarPublicacionOriginal		Y	(\)|\]|\?)$		N		N	N	N
5	lugarPublicacionOriginal		Y	(\(|\)		N		N	N	N
6	lugarPublicacionOriginal		Y	(\-[0-9]{1,2}\-[0-9]{1,2})$		N		N	N	N

Help | OK | Get fields | Cancel

Expresiones regulares en PDI

String operations

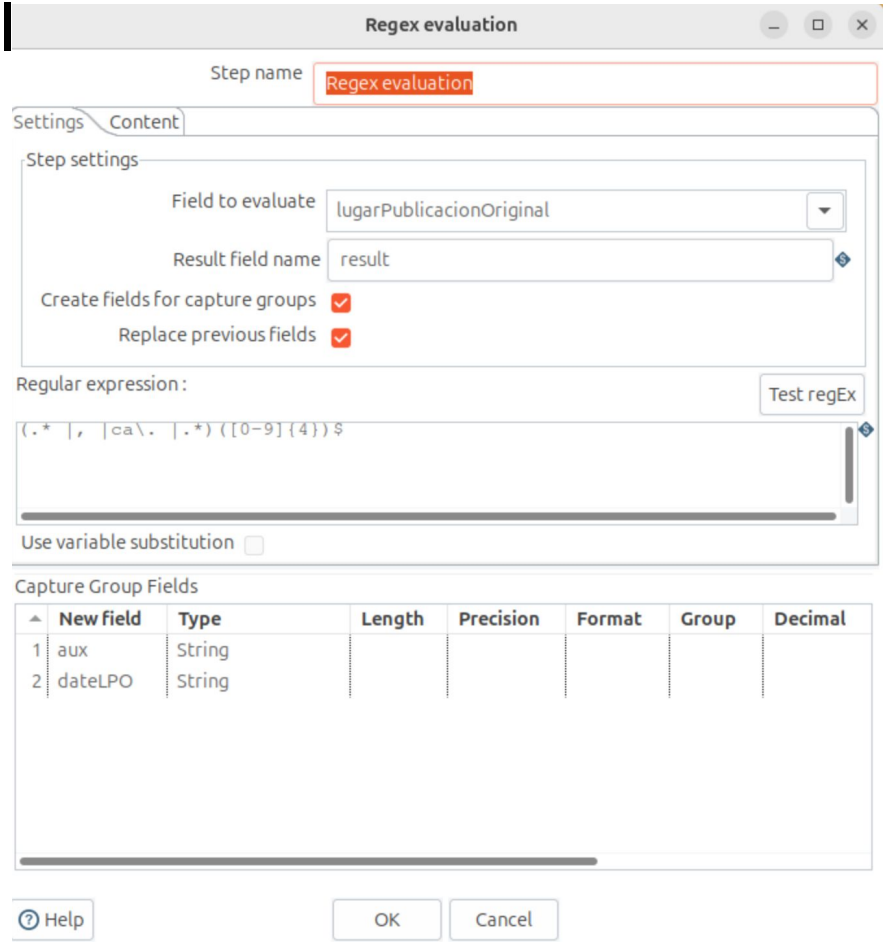


Expresiones regulares en PDI

Regex evaluation

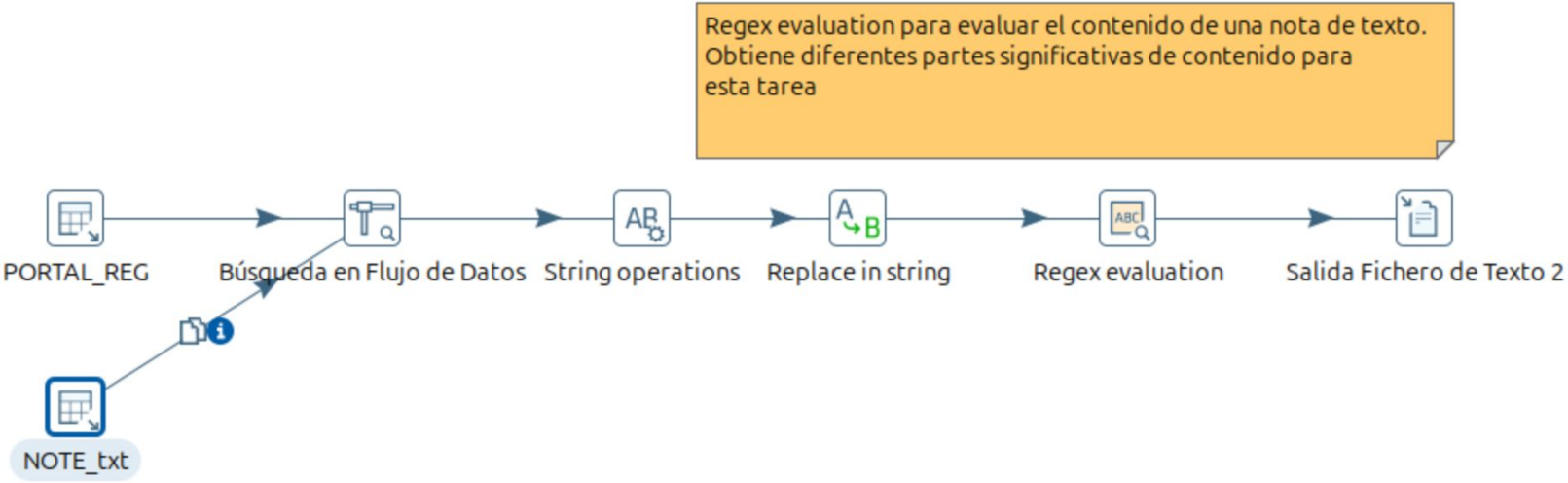
Extraer fecha de publicación de la nota de mención original de una obra.

Debe estar en formato YYYY



Expresiones regulares en PDI

Regex evaluation



Expresiones regulares en PDI

Regex evaluation

Step name: Regex evaluation

Settings | Content

Step settings

Field to evaluate: nota

Result field name: result

Create fields for capture groups ☑

Replace previous fields ☑

Regular expression:

Use variable substitution ☐

Test regEx

Capture Group Fields

#	New field	Type	Length	Precision	Format	Group	Decimal	Currency	Null If	Defaul	Trim
1	local	String									both
2	provincia	String									none

Help | OK | Cancel

Regular expression evaluation

Regular expression

Please enter a new regular expression or modify.

```
(.*, )?\(?(.+)\)?$
```

The regular expression was successfully compiled.

Values to test

Value1

Value2

Value3

Capture

Capture from value: una localización: Santiago de la Rivera, Murcia

Capture group fields (2: una localización: Santiago de la Rivera, Murcia

OK | Cancel

Expresiones regulares en PDI

(.*,)?\(?(.+)\)?$

Examine preview data

notaOriginal	result	local	provincia
Auxiliar localización: Madrid	Y	<null>	Madrid
Auxiliar localización: Salnés, Pontevedra	Y	Salnés,	Pontevedra
Auxiliar localización: Albaicín, Granada	Y	Albaicín,	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Albaicín, Granada	Y	Albaicín,	Granada
Auxiliar localización: Albaicín, Granada	Y	Albaicín,	Granada
Auxiliar localización: Albaicín, Granada	Y	Albaicín,	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Granada	Y	<null>	Granada
Auxiliar localización: Motril, Granada	Y	Motril,	Granada
Auxiliar localización: Sanlúcar de Barrameda, Cádiz	Y	Sanlúcar de Barrameda,	Cádiz
Auxiliar localización: A Coruña	Y	<null>	A Coruña
Auxiliar localización: Toledo	Y	<null>	Toledo
Auxiliar localización: Jaén	Y	<null>	Jaén
Auxiliar localización: Córdoba	Y	<null>	Córdoba
Auxiliar localización: Palencia	Y	<null>	Palencia
Auxiliar localización: Tuñón, Asturias	Y	Tuñón,	Asturias
Auxiliar localización: Olot, Girona	Y	Olot,	Girona

Expresiones regulares en PDI

Funciones y operadores de expresiones regulares para Mysql

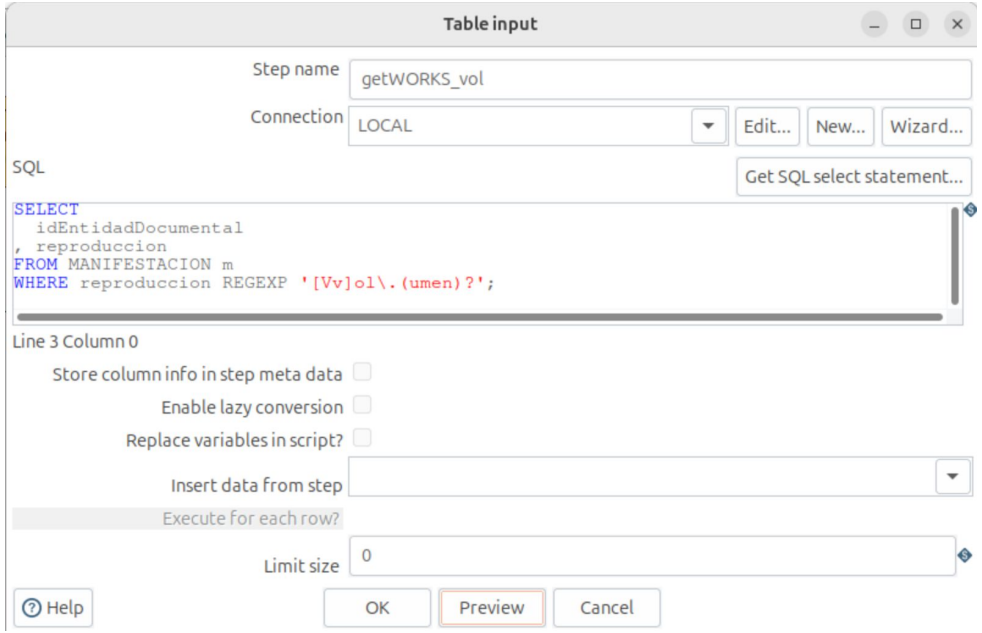
Name	Description
NOT REGEXP	Negation of REGEXP
REGEXP	Whether string matches regular expression
REGEXP_INSTR()	Starting index of substring matching regular expression
REGEXP_LIKE()	Whether string matches regular expression
REGEXP_REPLACE()	Replace substrings matching regular expression
REGEXP_SUBSTR()	Return substring matching regular expression
RLIKE	Whether string matches regular expression

Fuente: https://dev.mysql.com/doc/refman/8.4/en/regexp.html

Expresiones regulares en PDI

Uso REGEX de Mysql

```
SELECT

  idEntidadDocumental

, lugarPublicacionOriginal

, reproduccion

FROM MANIFESTACION m

WHERE reproduccion REGEXP '[Vv]ol\.(umen)?';
```



Expresiones regulares en PDI

Uso REGEX de Mysql

Resultado:

Edició digital basada en l' edició d'Alexandre Micha, Genève, Droz 1978-1983, **vols.**

Altra ed.: Catalan Review,**Volume** VI, numbers 1-2 (1992), pp. 393-399.

Otra ed.: Celestinesca,**vol.** 14, núm. 2 (nov. 1990), pp. 3-39

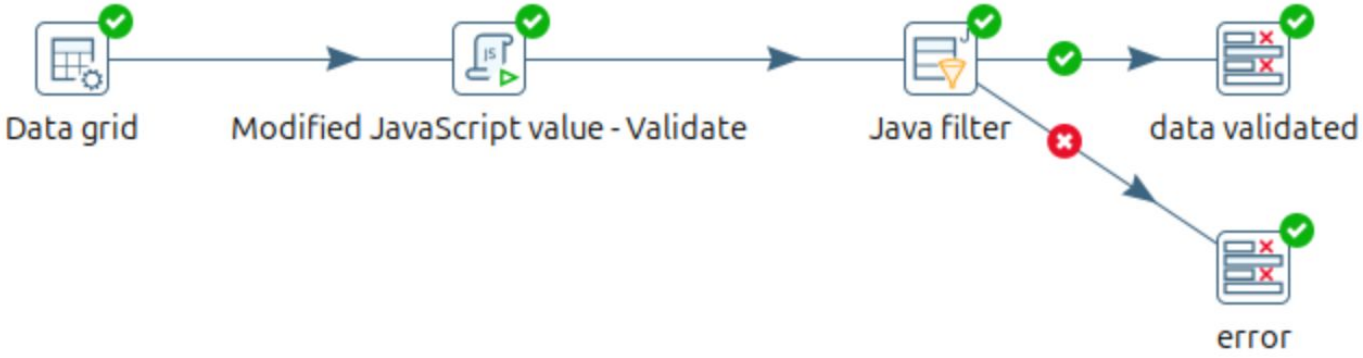
Edición digital a patir de la de Valencia, Herederos de Gerónimo Conejos, 1752, **vols.**

Altra ed.:Estudios Lulianos,**Vol.** 29 (1989), Fasc. 1, pp. 1-23. Fasc. 2, pp. 101-124

Expresiones regulares en PDI

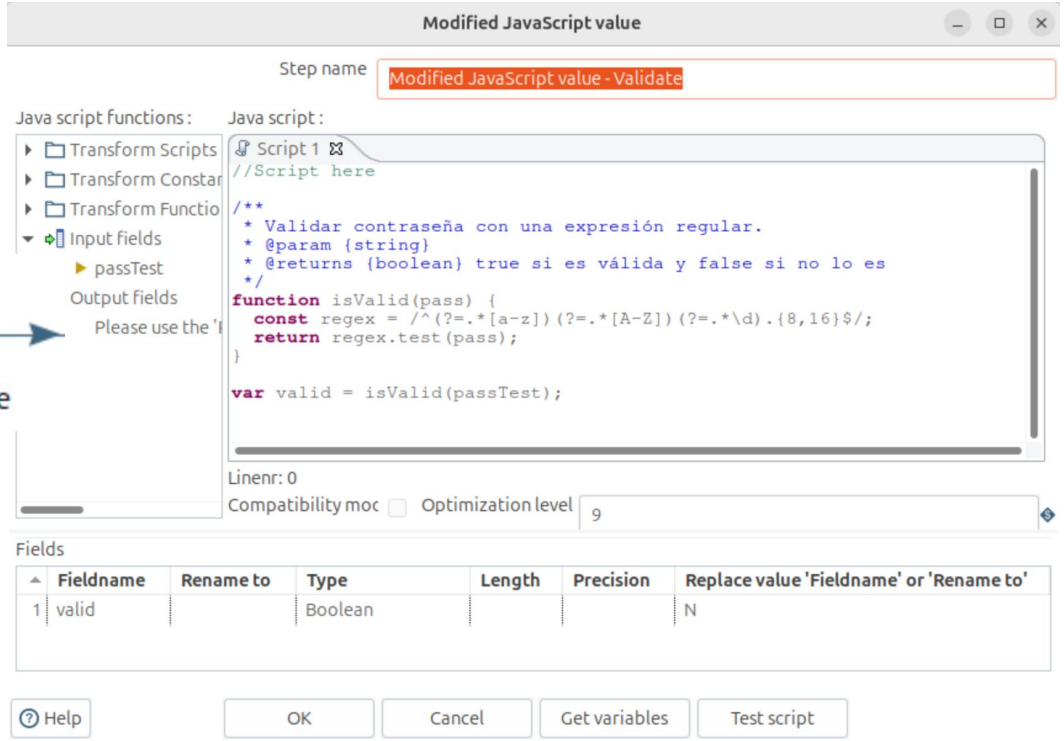
JavaScript

Variante del ejercicio 2 - Validar una contraseña mediante una expresión regular. La contraseña debe tener entre 8 y 16 caracteres y contener al menos una letra minúscula, una mayúscula y un número.



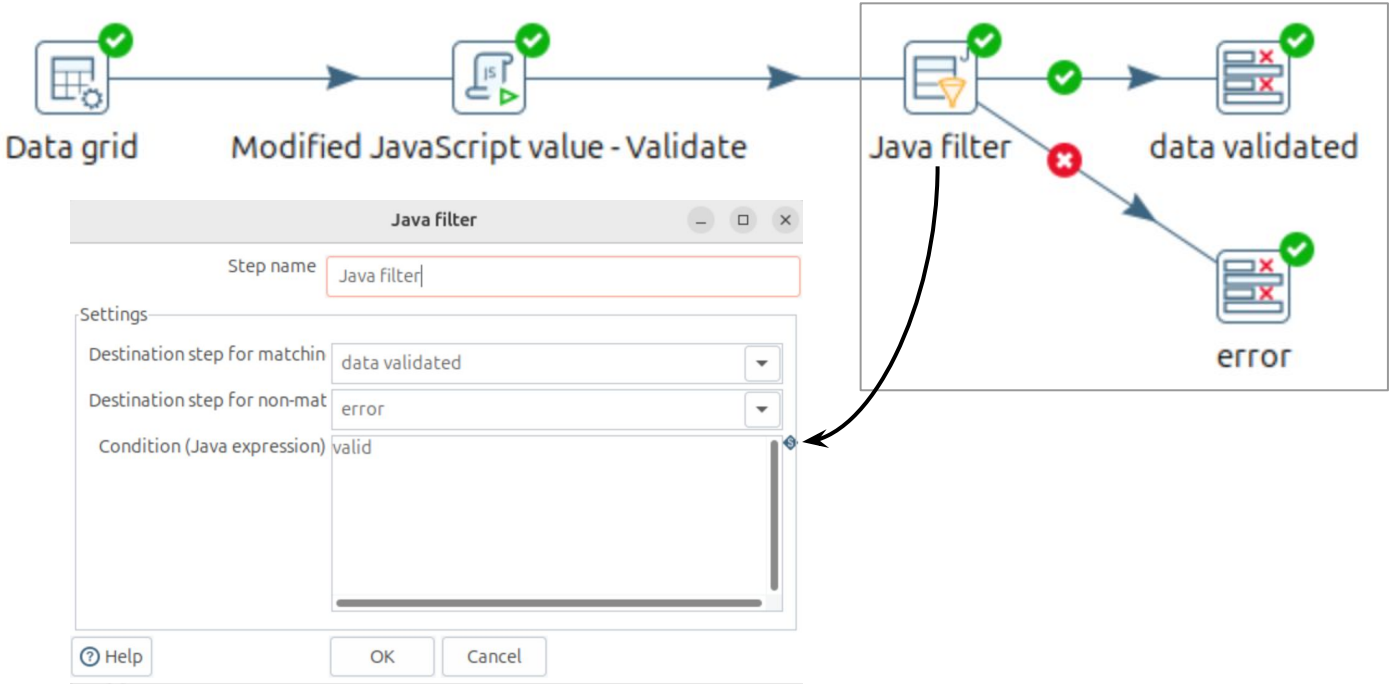
Expresiones regulares en PDI

Modified JavaScript value



Expresiones regulares en PDI

Java filter



Referencias

- Kelhini, F. K. (2024). Text processing with JavaScript : regular expressions, tools, and techniques for optimal performance / (First edition.). The Pragmatic Programmers, LLC.
- MySQL 8.4 Reference Manual,Regular Expressions dev.mysql.com/doc/refman/8.4/en/regexp.html
- Learn SQL SQL Server learn.microsoft.com/es-es/sql/relational-databases/regular-expressions/overview?view=sql-server-ver17
- Pentaho Data Integration Regex Evaluation pentaho-public.atlassian.net/wiki/spaces/EAI/pages/371558254/Regex+Evaluation (earlier version)

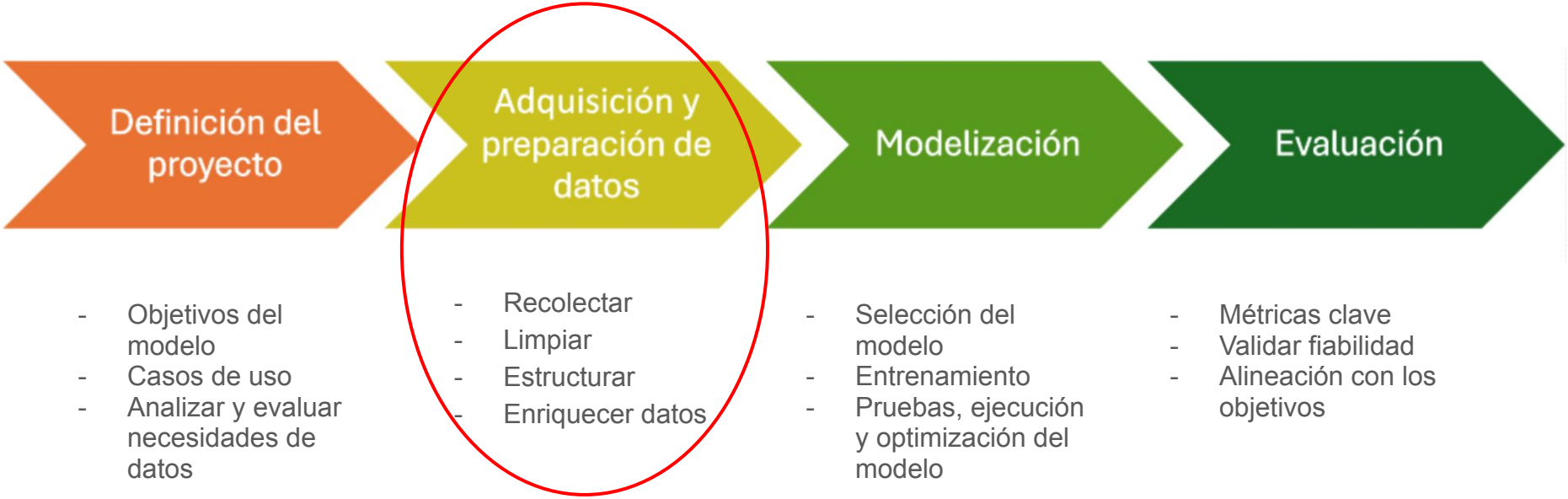
Procesos ETL

Tema 6: Limpieza, transformación y normalización de datos. Parte II

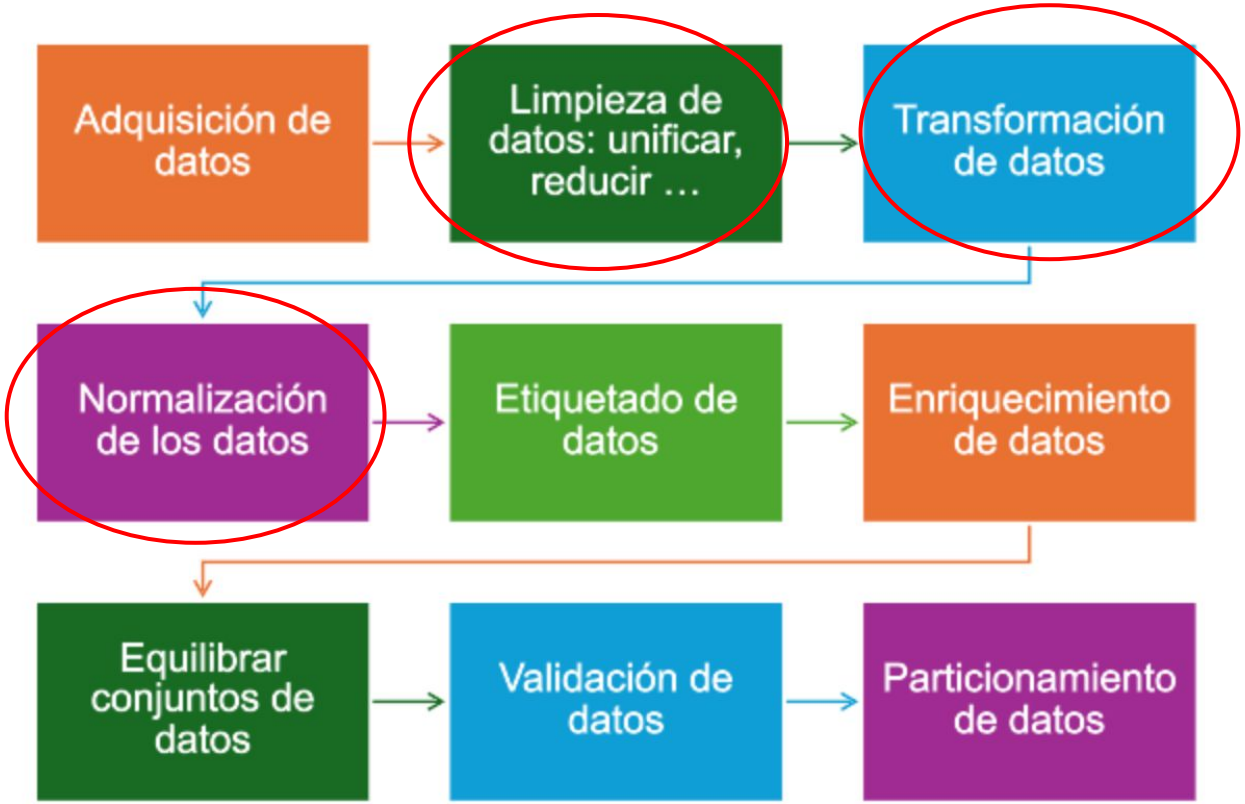
Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Ciclo de vida del proyecto IA



Preparación de datos



Limpieza, transformación y normalización

Limpieza de datos

Proceso esencial que elimina errores y anomalías en *raw data* con el objetivo de asegurar la **calidad** y **fiabilidad** de los datos, aspecto crítico para optimizar el rendimiento y la precisión de los modelos de IA.

Forma parte de la calidad de los datos y de la gobernanza de datos.

¿Dónde resolver los problemas con la calidad de los datos? ¿En la fuente?

➔ Diseñar transformaciones de limpieza de datos reutilizables

Implica entre otras tareas:

- **Corregir de errores, eliminar datos incorrectos o no válidos**
- **Gestionar los valores que faltan**
- **Eliminar o minimizar la redundancia de datos**

Corregir errores

Eliminar las inconsistencias, imprecisiones y errores tipográficos para asegurar la **calidad** y **uniformidad** de la información.

Errores tipográficos: fallos de escritura en la introducción de datos

"...cancionero catalán de la universidad de zAragoza"

"Edició digital a partir de l'edició de Mariano Baselga y Ramírez, Zaragoza, Cecilio Gasca, 18963".

"El espejo de la veredad"

&#0    

"Edición digital basada ↲↲↲en la de ParíS, Sociedad de Ediciones Literarias y Artísticas, [s.a.]."

Corregir errores

Errores de formato: diferentes representaciones de la misma información

aprox 1581-1639 → aproximadamente 1581-1639

Siglo XVII →S. 17º

1609-12-9 →1609-12-09 (formato YYYY-MM-DD)

1810/09/08 →1810-09-08 (formato YYYY-MM-DD)

02/12/2000 15-00-24 →2000-12-02 (formato YYYY-MM-DD)

12 de diciembre 1925 →1925-12-12 (formato YYYY-MM-DD)

chi / zho, zho, zh, chino, chinés →chi

spa, es, esp, español → es

mapa, material cartográfico (impreso) →cartográfico

EE.UU., USA, E.E.U.U., EUA → EE.UU.

Corregir errores

Inconsistencias: datos contradictorios, incorrectos o no válidos para el dominio

“Edició digital a partir de l'edició de Mariano Baselga y Ramírez, zAragoza, Cecilio Gasca, 18963”.

Fechas de un autor (nacimiento-fallecimiento) 1733-1921 → vivió 188 años??

Fechas de un autor (nacimiento-fallecimiento) 1560-1536 → invertidas, error tipográfico??

Errores en códigos de ISBN o ISSN 978-84-690-2311-02 → ISBN incorrecto

El mismo autor aparece como: "Federico García Lorca", "Lorca F.", "García, Federico"

"Presentación contextualizada de cartas enviadas por Gómez de Baquero a Miguel de Unamuno, además de una dirigida a Ortega y Gasset. La correspondencia entre aquellos dos autores, iniciada por Unamuno en 1896, permite el seguimiento de la evolución…"

Ejemplar -5 páginas

Valores faltantes

Los valores faltantes es un problema común en la adquisición de datos y pueden ser debido a errores, datos sensibles, no suministrados, etc.

Gestionar valores que faltan por **reemplazo, interpolación o eliminación**.

- Registros bibliográficos sin título <dc:title>null</dc:title> → se registra el error y se descarta registro
- Registros bibliográficos sin autor <dc:creator>null</dc:creator> → se deja vacío o normaliza a “no dado” para el análisis, no normaliza a “Anónimo”
- Género del autor o autora faltante → normalizar a “Desconocido”
- Formato de una obra <dc:format></dc:format> → normalizar a “sin definir”

Redundancia de datos

El Objetivo principal es es garantizar datos únicos.

Reducir el tamaño del dataset y optimizar el entrenamiento del modelo.

Duplicados completas → eliminar duplicado

Dos registros bibliográficos con mismo título y mismo autor pero uno es está en PDF y otro en EPUB → fusionar y consolidar los campos más completos reflejando la relación.

El mismo autor aparece con distintos nombre: "Azorín” y “José Martínez Ruiz” → reflejar alternativas válidas que identifican al autor.

Materia o tema de un libro que proviene de diferentes fuentes Universidades -- España -- Siglo-19 y Universidades -- España -- S-XIX → normalizar y eliminar duplicados

Estrategias y técnicas

- Uso de expresiones regulares para encontrar y reemplazar formatos específicos.
- Limpieza para eliminar espacios en blanco, caracteres especiales, puntuación no deseada o caracteres extraños.
- Algoritmos para la comparación de cadenas (*string matching*), se identifican errores tipográficos, duplicados.
- Validación con datos del dominio →Mapeo de datos a vocabularios controlados.
- Estandarización aplicando un formato único o valor de referencia.

Dirección en “bruto”

"Avda. de la Estación, s/n, entresuelo C - 03003 (ALICANTE)"

¿Qué tareas o pasos pensáis que deberíamos hacer para “limpiar y normalizar”?

Ejemplo

1- Normalizar nombres

"Avda. de la Estación, s/n, entresuelo C - 03003 (ALICANTE)"

La abreviatura Avda. debe ser estandarizado a "**Avenida**"

Se debe determinar que hacemos con "s/n" podría convertirse en "sin número" o eliminarlo si no es necesario

Ejemplo

2- Separar en los distintos componentes

"Avda. de la Estación, s/n, entresuelo C - 03003 (ALICANTE)"

Tipo de vía → Vocabulario controlado

Nombre de la calle/vía

Identificador

Información adicional a la dirección: piso y puerta

CP, que ya nos da la información que es de Alicante

Ejemplo

3- Eliminar redundancias

El CP ya indica que es Alicante. El nombre de la localidad “(ALICANTE)" es redundante y se podría eliminar o estandarizar.

¿Es correcta la dirección postal?

Comprobar datos por ejemplo consultado bases de conocimiento o APIs disponibles.

- Postcode www.postcode.eu/es www.postcode.eu/es/products/validate-api
- Cartociudad datos.gob.es/es/aplicaciones/direcciones-postales-de-cartociudad-espana plataforma.idee.es/cnig-api

Este proceso de limpieza y normalización asegura que los datos sean coherentes, precisos y fáciles de usar:

Tipo Vía	Nombre Vía	Número	Piso/Puerta	Código Postal	Localidad
Avenida	Estación	s/n	Entresuelo C	03003	Alicante

Limpieza de datos

- No hay un único paso de limpieza de datos
- El proceso de depuración de datos comienza en el mismo paso de Extracción
  - Muchos de los pasos de entrada contienen funciones para leer los datos en un formato específico
- Generalmente es mejor leer los datos sin procesar “tal cual” están en origen (*raw data*) y aplicar en el ETL los pasos de limpieza necesarios
  - Por ejemplo en el paso Table input se puede aplicar cierta limpieza de datos desde las sentencias SQL. → Se dificulta la auditoría
- Se facilita la definición de flujos de error o procesamiento condicional para redirigir los datos que no satisfacen los requisitos.

Transformación y normalización

El objetivo de la **transformación** es preparar los datos para que puedan ser utilizarlos directamente.

Las transformaciones pueden incluir tareas como:

- Seleccionar campos, filtrar filas, ordenar, agrupar y aplicar funciones de agregación
- Realizar fusión de varias fuentes de datos
- *Feature Engineering*, transformación y creación de nuevas características a partir de datos existentes para mejorar la precisión y el rendimiento de los modelos
- Transformar datos no estructurados

Transformación y normalización

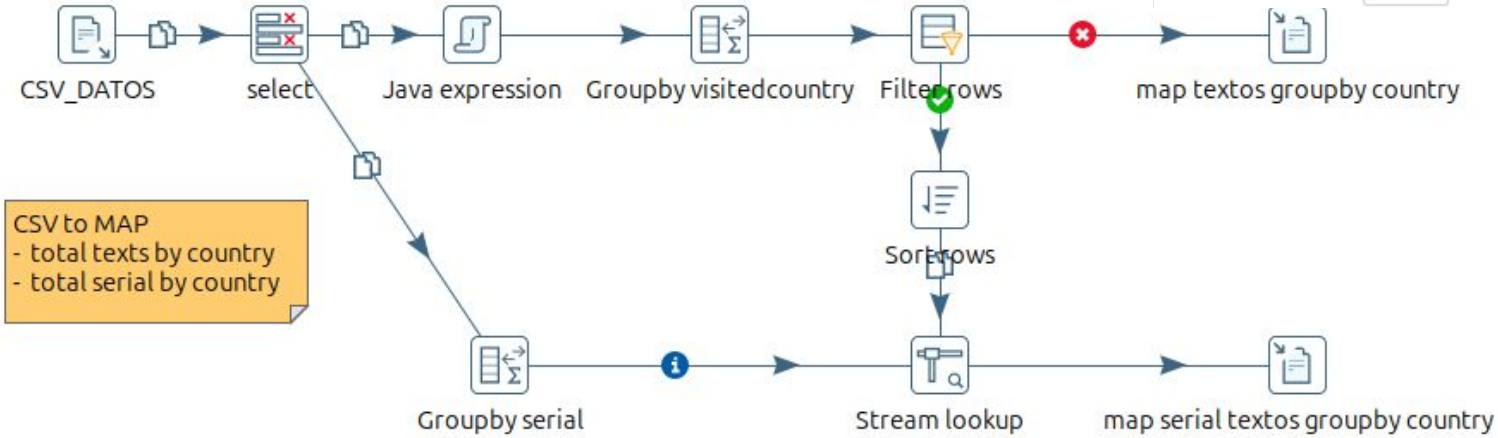
Ejemplo de transformación con **selección** de campos, **filtrado**, **ordenación**, **agrupación** y **agregación**

Examine preview data

Rows of step: Groupby serial (27 rows)

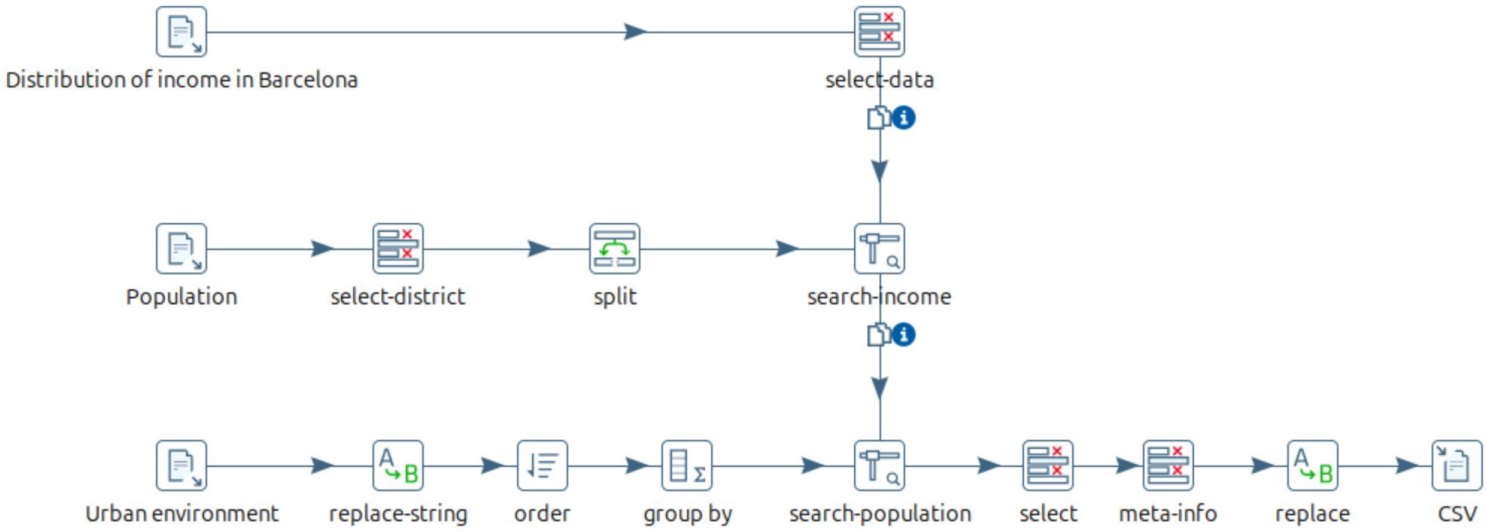
	PAÍSES_VISITADOS	countTextos	countSerial
1	<null>	4	2
2	Paraguay	21	13
3	Santo Domingo	7	5
4	varios	261	26
5	Martinica	5	4
6	Guyana Francesa	63	3
7	Argentina	182	26
8	Chile	119	12
9	Perú	94	14
10	Bolivia	26	7

Close



Transformación y normalización

Ejemplo de **fusión** de varias fuentes de datos



Feature Engineering

Proceso de transformar los datos sin procesar en nuevas variables, características o *features*, para mejorar el rendimiento de los modelos.

Representan **aspectos esenciales** de los datos, **relevantes** para el problema planteado.

Las características son variables clave de los datos y pueden ser textuales, numéricas o categóricas.

¿Por qué es importante?

- Mejorar la precisión y la eficiencia
- Mejora la interpretabilidad del modelo
- Puede reducir el sobreajuste

Feature Engineering

- **Creación** de nuevas características: combinar, transformar, extraer o derivar nuevas, por ejemplo *Splitting*
- **Codificación** de variables categóricas, por ejemplo *One-Hot Encoding*
- **Transformación** variables numéricas continuas en categorías, *Binning*
- **Enriquecimiento** de datos
- **Escalado** de datos
- Tratamiento de los ***Outliers***

Feature Engineering

Ejemplo de **creación de nuevas características** mediante **transformación** y **combinación** de datos

```
//MENCION DE PUBLICACION ORIGINAL
var lpo = place_of_publication+ publisher + editor + date_created;
if(place_of_publication != null)
    lpo = place_of_publication
if(publisher != null)
    lpo = lpo + ", " + publisher;

if(date_created != null)
    lpo = lpo + ", " + date_created;

//Detalles físicos mapeo directo de campos
//dimensions extension digital_file_characteristics
//var characteristicsAux = digital_file_characteristics
```

Feature Engineering

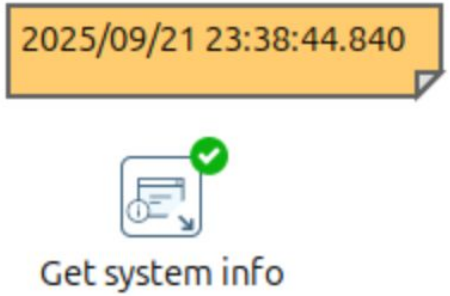
Ejemplo de **creación de nuevas características** mediante **extracción**

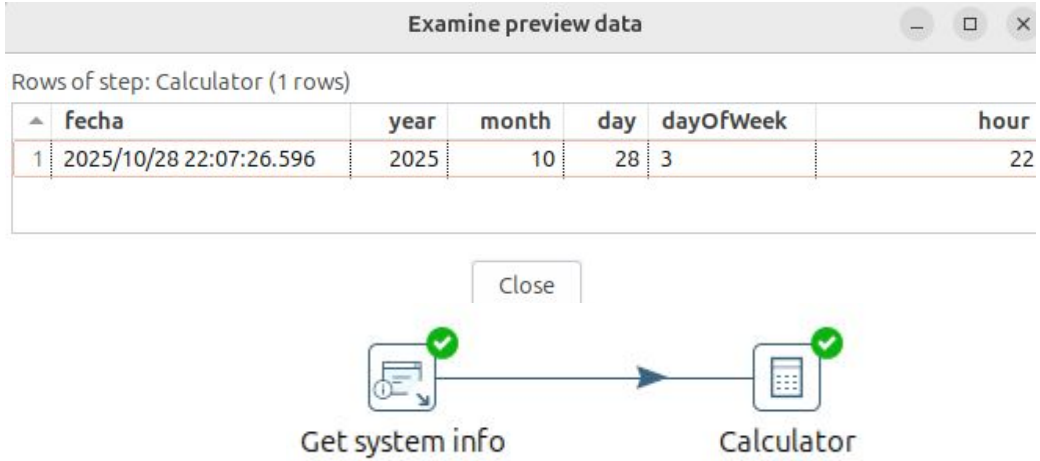
```
import pandas as pd

# fechas datetime
data = pd.DataFrame({'Fecha': pd.to_datetime(['2025-09-08', '2024-06-15', '2025-09-20'])})

data['year'] = data['Fecha'].dt.year
data['month'] = data['Fecha'].dt.month
data['day'] = data['Fecha'].dt.day
data['day_week'] = data['Fecha'].dt.dayofweek  #Lunes=0, Domingo=6
print(data)
```

	Fecha	year	month	day	day_week
0	2025-09-08	2025	9	8	0
1	2024-06-15	2024	6	15	5
2	2025-09-20	2025	9	20	5

Limpieza, transformación y normalización

Ejemplo de **creación de nuevas características** mediante **extracción**





	fecha	year	month	day	dayOfWeek	hour
1	2025/10/28 22:07:26.596	2025	10	28	3	22

Feature Engineering

Ejemplo de c**reación de nuevas características** mediante **splitting**

"Avda. de la Estación, s/n, entresuelo C - 03003 (ALICANTE)"

Tipo Vía	Nombre Vía	Número	Piso/Puerta	Código Postal	Localidad
Avenida	Estación	s/n	Entresuelo C	03003	Alicante

Feature Engineering

Ejemplo de **creación de nuevas características** mediante ***splitting***

```python
import pandas as pd

data = {'Dirección completa': [
'Alonso Cano 26, Alicante, 03012', 'Calle Sierpes 56, Sevilla, 41004']}
df = pd.DataFrame(data)

df[['Calle', 'Ciudad', 'CP']] = df['Dirección completa'].str.extract(
    r'([\w\s]+\b[0-9]+\b),\s([\w\s]+),\s(\b\d{5}\b)')

print(df)
```

```
                 Dirección completa             Calle    Ciudad     CP
0   Alonso Cano 26, Alicante, 03012    Alonso Cano 26  Alicante  03012
1  Calle Sierpes 56, Sevilla, 41004  Calle Sierpes 56   Sevilla  41004
```

Feature Engineering

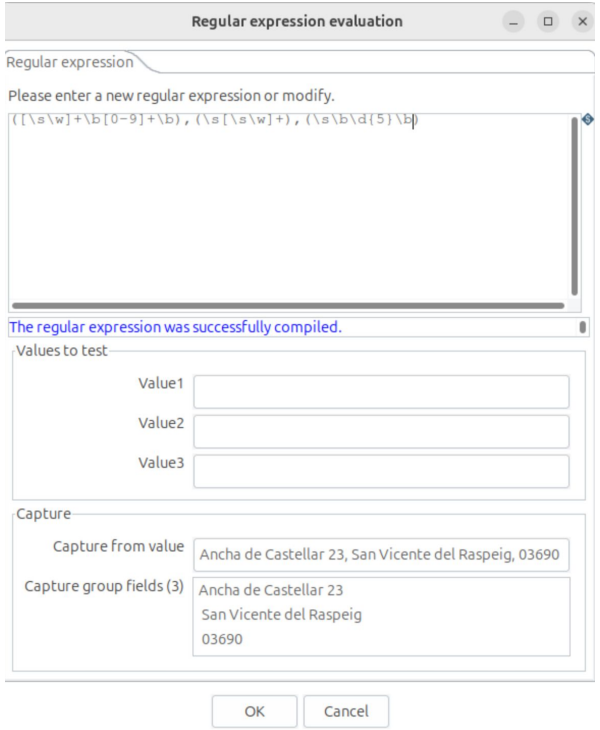
Ejemplo de **creación de nuevas características** mediante ***splitting***

```
import pandas as pd

data = {'Full_Address': [
    '123 Elm St, Springfield, 12345', '456 Oak Rd, Shelbyville, 67890']}
df = pd.DataFrame(data)

df[['Street', 'City', 'Zipcode']] = df['Full_Address'].str.extract(
    r'([0-9]+\s[\w\s]+),\s([\w\s]+),\s(\d+)')

print(df)
```

```
                     Full_Address      Street         City Zipcode
0  123 Elm St, Springfield, 12345  123 Elm St  Springfield   12345
1  456 Oak Rd, Shelbyville, 67890  456 Oak Rd  Shelbyville   67890
```

Feature Engineering

Ejemplo de transformación y creación de nuevas características - **Splitting**



Feature Engineering

$a Poesía infantil y juvenil $x Historia y crítica

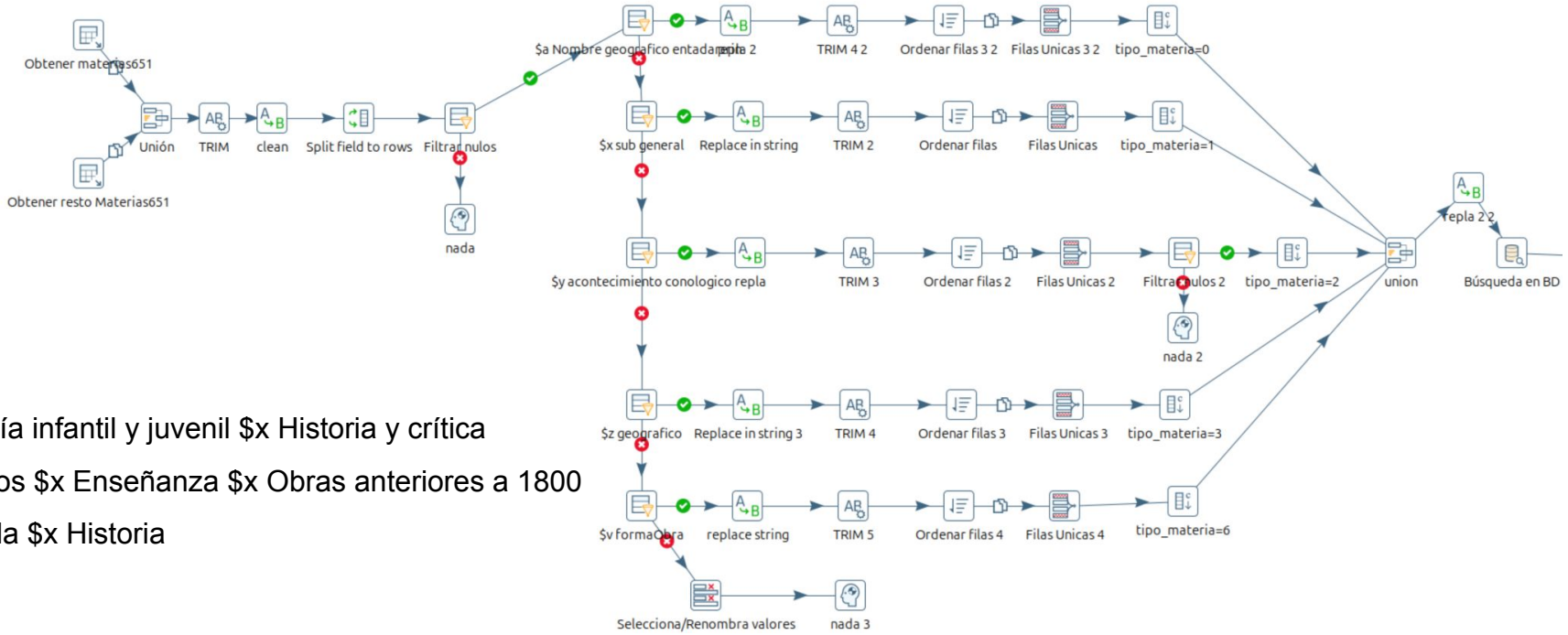
$a Sordos $x Enseñanza $x Obras anteriores a 1800

$a Sevilla $x Historia

LC Control Number: sh 85083794
HEADING: Mercury

```
000 00558cz a2200217n 450
001 4734282
005 19900221112154.6
008 860211i| anannbabn |a ana
035 __ |a (DLC)sh 85083794
906 __ |t 8528 |u fk03 |v 0
010 __ |a sh 85083794
040 __ |a DLC |c DLC |d DLC
053 _0 |a QD181.H6 |c Chemistry
053 _0 |a TA480.M4 |c Engineering materials
053 _0 |a TN271.M4 |c Prospecting
053 _0 |a TP245.M5 |c Chemical technology
150 __ |a Mercury
450 __ |a Hydrargyrum
450 __ |a Quicksilver
550 __ |w g |a Liquid metals
953 __ |a xx00 |b fg07
```

Feature Engineering



$a Poesía infantil y juvenil $x Historia y crítica

$a Sordos $x Enseñanza $x Obras anteriores a 1800

$a Sevilla $x Historia

Feature Engineering

Codificación de variables categóricas

One-Hot Encoding

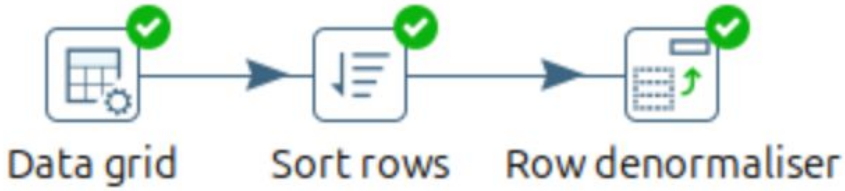
Convierte las variables categóricas en indicadores binarios

[15]:
```
import pandas as pd

data = {'ContentType': ['audio', 'video', 'graphic', 'text', 'online']}
df = pd.DataFrame(data)

df_encoded = pd.get_dummies(df, columns=['ContentType'], prefix='content')

print(df_encoded)
```

```
   content_audio  content_graphic  content_online  content_text  content_video
0           True            False           False         False          False
1          False            False           False         False           True
2          False             True           False         False          False
3          False            False           False          True          False
4          False            False            True         False          False
```

Feature Engineering

Codificación de variables categóricas



Feature Engineering

Binning, transformación de variables continuas discretas en categorías

Binning

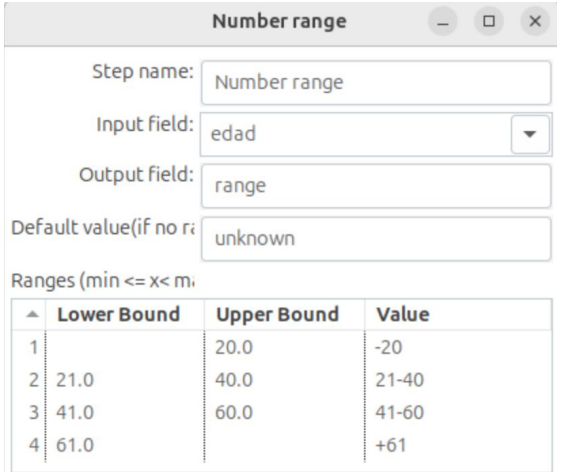
Divide los datos originales en intervalos (bins), y después se reemplazan por un valor general calculado para ese bin Efecto suavizado de datos y además, puede reducir el riesgo de sobreajuste en conjuntos de datos pequeños

```python
import pandas as pd

data = {'edad': [23, 45, 18, 34, 67, 50, 21]}
df = pd.DataFrame(data)

bins = [0, 20, 40, 60, 120]
labels = ['0-20', '21-40', '41-60', '+61']

df['Grupo_edad'] = pd.cut(df['edad'], bins=bins, labels=labels, right=False)

print(df)
```

```
   edad Grupo_edad
0    23      21-40
1    45      41-60
2    18       0-20
3    34      21-40
4    67        +61
5    50      41-60
6    21      21-40
```

Feature Engineering

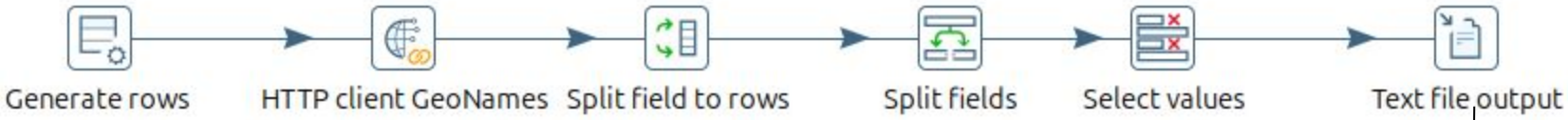
Binning, transformación de variables continuas discretas en categorías



	nombre	edad	range
1	Pepe	45	41-60
2	Ana	52	41-60
3	Eva	25	21-40
4	Jorge	23	21-40
5	Alejandro	78	+61
6	Pablo	65	+61

Feature Engineering

Enriquecimiento de datos, conectar con datos externos para aumentar el valor de nuestro conjunto de datos



Rows of step: Select values (255 rows)

	f1	f2	f3	f4	f5	f6	f7	f8	f9	f10
1	iso alpha2	iso alpha3	iso numeric	fips code	name	capital	areaInSqKm	population	continent	languages
2	AD	AND	020	AN	Andorra	Andorra la Vella	468.0	77006	EU	ca
3	AE	ARE	784	AE	United Arab Emirates	Abu Dhabi	82880.0	9630959	AS	ar-AE,fa,en,hi,ur
4	AF	AFG	004	AF	Afghanistan	Kabul	647500.0	37172386	AS	fa-AF,ps,uz-AF,tk
5	AG	ATG	028	AC	Antigua and Barbuda	St John's	443.0	96286	NA	en-AG
6	AI	AIA	660	AV	Anguilla	The Valley	102.0	13254	NA	en-AI

Feature Engineering

Enriquecimiento de datos

OpenWeather

https://api.openweathermap.org/data/2.5/weather?q=Zaragoza&APPID=


```
{
  "coord": {
    "lon": -0.8773,
    "lat": 41.6561
  },
  "weather": [
    {
      "id": 800,
      "main": "Clear",
      "description": "clear sky",
      "icon": "01d"
    }
  ],
  "base": "stations",
  "main": {
    "temp": 302.75,
    "feels_like": 301,
    "temp_min": 302.4,
    "temp_max": 303.16,
    "pressure": 1014,
    "humidity": 17,
    "sea_level": 1014,
    "grnd_level": 976
  },
  "visibility": 10000,
  "wind": {
    "speed": 6.69,
    "deg": 310
  },
  "clouds": {
    "all": 0
  },
  "dt": 1757006653,
  "sys": {
    "type": 2,
    "id": 2003310,
    "country": "ES",
    "sunrise": 1756963963,
    "sunset": 1757010766
  },
  "timezone": 7200,
  "id": 3104324,
  "name": "Zaragoza",
  "cod": 200
}
```

Transformación y normalización

El objetivo de la **transformación** es preparar los datos para que puedan ser utilizarlos directamente.

Las transformaciones pueden incluir tareas como:

- Seleccionar campos, filtrar filas, ordenar, agrupar y aplicar funciones de agregación
- Realizar fusión de varias fuentes de datos
- *Feature Engineering*, transformación y creación de nuevas características a partir de datos existentes para mejorar la precisión y el rendimiento de los modelos
- **Transformar datos no estructurados**

Clasificación de los datos según formato y esquema

	Datos Estructurados	**Datos No Estructurados**	**Datos Semiestructurados**
Estructura	Modelo de datos predefinido y estricto	Sin modelo de datos predefinido	Estructura más flexible y jerárquica Uso de metadatos
Almacenamiento	Formato tabular Esquema rígido bases de datos relacionales y almacenes de datos*	Formato nativo, bases de datos NoSql, data lakes	Bases de datos NoSQL
Análisis	Fácil y rápido	Más complejo y requieren habilidades y herramientas especializadas	Moderadamente complejo
Ejemplos	Tablas, bases de datos Sql, hojas de cálculo, csv*	Textos, imágenes, videos, multimedia, IoT, social media	JSON, XML
Casos de uso	Machine learning (ML)	PLN Modelos IA generativa RAG (Generación Aumentada por Recuperación) Análisis predictivos	

Transformar datos no estructurados

Datos no estructurados : no tiene un formato predefinido o un modelo de datos fijo ni estandarizado.

PDFs, archivos multimedia (audios, imágenes y vídeos), contenido en redes sociales, logs de servidor, datos IoT, correos, etc.

Según Gartner entre el 80 y el 90 % son datos no estructurados.

Muy valiosos, pero difíciles de procesar directamente con herramientas de análisis tradicionales, especialmente para IA.

La información está ahí, pero hay que 'enseñarle' a los modelos a verla.

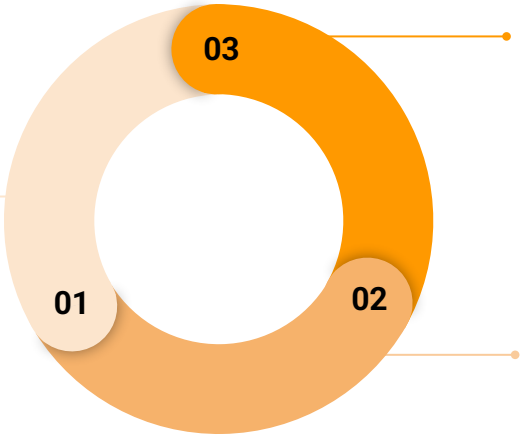
Transformar datos no estructurados

El Desafío: preparar los datos no estructurados para la IA

***Raw data* directamente de la fuente**
Diversas fuentes

Textos **sin procesar,** sin limpiar, imágenes sin etiquetar, audios sin transcripción

Es más complejo que los modelos aprendan de esto



Modelo IA

Se genera conocimiento

Preprocesamiento de los datos

El proceso de transformación: **limpieza y estructuración** que hace que el dato sea adecuado y útil para el entrenamiento de modelos

Transformar datos no estructurados

¿Cómo?

- Preprocesamiento texto, puede implicar tareas como eliminar *stop-words, stemming* y *vectorizing* en los datos de texto para prepararlos para los modelos
- Transformación de Imágenes y video
- Transformación de Audio

El objetivo es convertir el dato en una representación numérica (**vector**), que capture su significado.

Transformar datos no estructurados

```python
import spacy

nlp = spacy.load("es_core_news_md")
text = nlp("""Los veía todas las mañanas durante esos días otoñales en que el cielo de Madrid finge sonrisas
    de primavera; sentados frente al borde de las aceras, la larga fila de vendedores ambulantes se extendía
    a lo largo de la calle.""")

# Filter stopwords using spaCy
refinedWords = [token.text for token in text if not token.is_stop]
print("Sin stop words:", refinedWords)
```

```
Sin stop words: ['veía', 'mañanas', 'otoñales', 'cielo', 'Madrid', 'finge', 'sonrisas', '\n    ', 'primavera',
';', 'sentados', 'frente', 'borde', 'aceras', ',', 'larga', 'fila', 'vendedores', 'ambulantes', 'extendía', '\n
', 'calle', '.']
```

```python
for word in text:
    print(word.text + ' --> ' + word.lemma_, '[' + word.pos_ + ']')
```

```
Los --> él [PRON]
veía --> ver [VERB]
todas --> todo [DET]
las --> el [DET]
mañanas --> mañana [NOUN]
durante --> durante [ADP]
esos --> ese [DET]
días --> día [NOUN]
otoñales --> otoñal [ADJ]
```

```
vendedores --> vendedor [NOUN]
ambulantes --> ambulante [ADJ]
se --> él [PRON]
extendía --> extender [VERB]

 la --> el [DET]
 calle --> calle [NOUN]
 . --> . [PUNCT]
```

Transformar datos no estructurados

```python
# Delete Stop-words using nltk

import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize

# Text unstructured data
text = "Los veía todas las mañanas durante esos días otoñales en que el cielo de Madrid finge sonrisas de primavera; sentados frente al bor

# Get Spanish stopwords and tokenize
liststopwords = set(stopwords.words('spanish'))
tokens = word_tokenize(text.lower())

# Remove stopwords
refinedTokens = [word for word in tokens if word not in liststopwords]

# print("Original:", tokens)
print(refinedTokens)
```

```
['veía', 'todas', 'mañanas', 'días', 'otoñales', 'cielo', 'madrid', 'finge', 'sonrisas', 'primavera', ';', 'sentados', 'frente', 'borde',
'aceras', ',', 'larga', 'fila', 'vendedores', 'ambulantes', 'extendía', 'largo', 'calle', '.']
```

```python
from nltk.stem import PorterStemmer

porterStemmer = PorterStemmer()
stemmed_words = [porterStemmer.stem(word) for word in refinedWords]
print("Stemmed words:", stemmed_words)
```

```
Stemmed words: ['veía', 'mañana', 'otoñal', 'cielo', 'madrid', 'fing', 'sonrisa', 'primavera', ';', 'sentado', 'frent', 'bord', 'acera',
',', 'larga', 'fila', 'vendedor', 'ambulant', 'extendía', 'call', '.']
```

Transformar datos no estructurados

Ejemplo *Vectorizing*

```python
# Bag of Words (BoW)
from sklearn.feature_extraction.text import CountVectorizer

text = ["Los veía todas las mañanas durante esos días otoñales en que el cielo de Madrid finge sonrisas de primavera;",
        "sentados frente al borde de las aceras, la larga fila de vendedores ambulantes se extendía a lo largo de la calle."
       ]

vectorizer = CountVectorizer()
X = vectorizer.fit_transform(text)

print(X.toarray())
print(vectorizer.get_feature_names_out())
```

```
[[0 0 0 0 0 1 2 1 1 1 1 1 0 0 1 0 0 0 0 1 0 1 1 1 1 1 1 0 0 1 1 0 1]
 [1 1 1 1 1 0 3 0 0 0 0 0 1 1 0 1 2 1 1 1 1 0 0 0 0 0 0 1 1 0 0 1 0]]
['aceras' 'al' 'ambulantes' 'borde' 'calle' 'cielo' 'de' 'durante' 'días'
 'el' 'en' 'esos' 'extendía' 'fila' 'finge' 'frente' 'la' 'larga' 'largo'
 'las' 'lo' 'los' 'madrid' 'mañanas' 'otoñales' 'primavera' 'que' 'se'
 'sentados' 'sonrisas' 'todas' 'vendedores' 'veía']
```

Procesos ETL

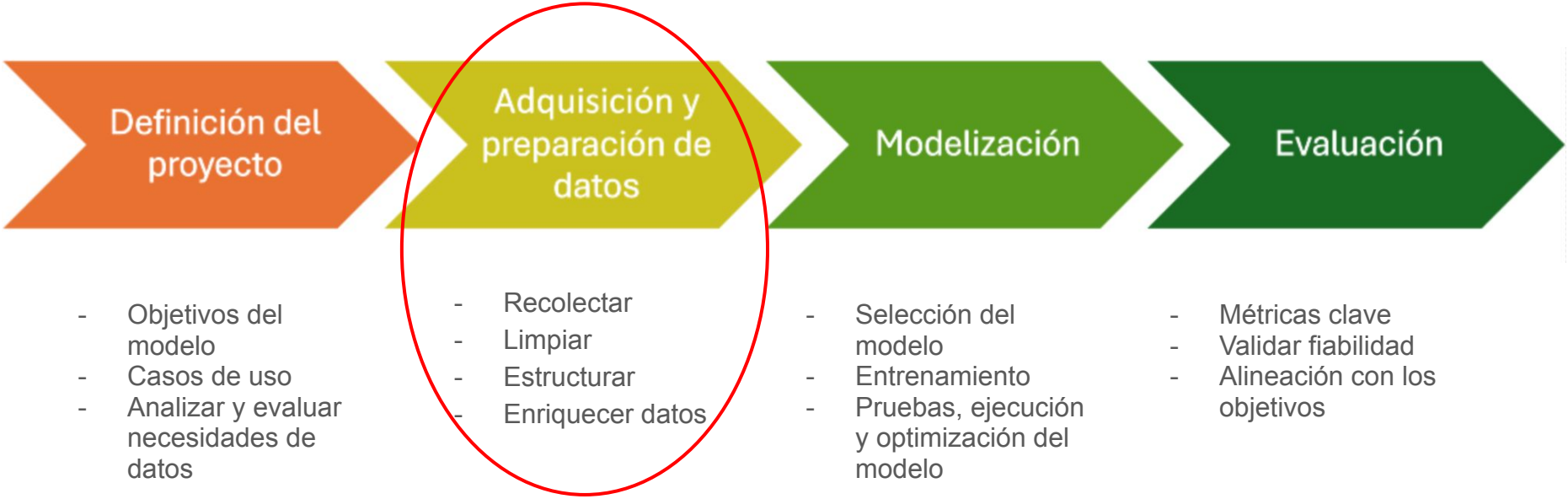
Tema 6: Limpieza, transformación y normalización de datos. Parte III

Adquisición y preparación de datos

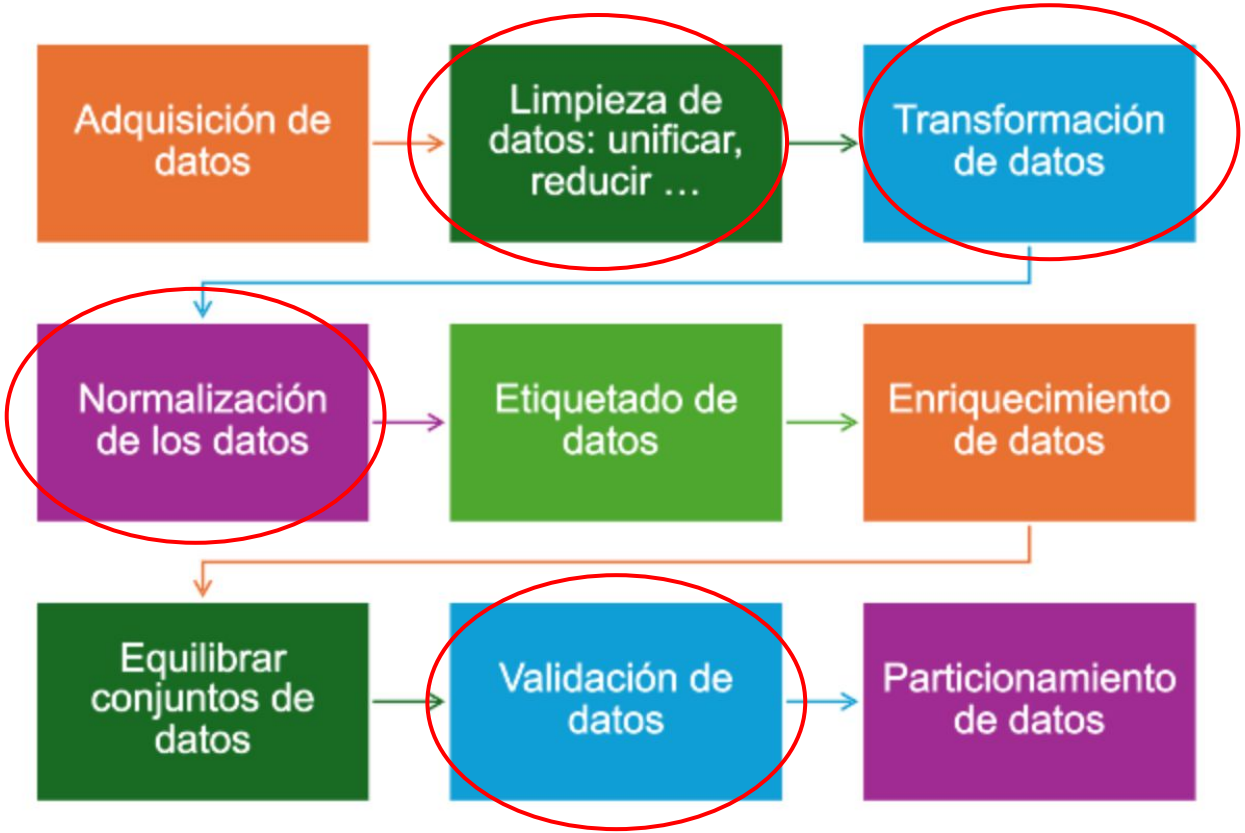
IA EPS UA

Universitat d'Alacant
Universidad de Alicante

Ciclo de vida del proyecto IA



Preparación de datos



Transformación y normalización

El objetivo de la **normalización** es escalar las características de un conjunto de datos a un rango común, asegurando que ninguna variable domine a otras debido a su magnitud.

Esto es crucial para mejorar la **equidad** y el **rendimiento** de los modelos.

Métodos de normalización:

- ***Min-Max Scaling***: escala para ajustarlos a un rango normalmente [0,1]
- ***Z-score*** *Standard Scaler*: ajusta los datos para que tengan una media de cero y una desviación estándar de uno

Gestión e identificación de valores atípicos (**Outliers**), son valores que se desvían significativamente de la mayoría de los datos.

Normalización de datos

Normalización de datos *Min-Max Scaling*

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

```
import numpy as np

edad = np.array([5, 16, 18, 15, 25, 35, 50])

print('Min-Max scaling: ',np.round((edad - edad.min())/(edad.max() - edad.min()),2))


Min-Max scaling:  [0.   0.24 0.29 0.22 0.44 0.67 1.  ]
```

Normalización de datos

Z-score (Standard Scaler)

$$x' = \frac{x - \bar{x}}{\sigma}$$

```
import numpy as np
edad = np.array([5, 16, 18, 15, 25, 35, 50])
print('Z-score: ',(edad - edad.mean())/edad.std())
print('Z-score: ',np.round(((edad - edad.mean())/edad.std()), 3))


Z-score:  [-1.33308858 -0.53736904 -0.39269276 -0.60970718  0.11367422  0.83705562
1.92212772]

Z-score:  [-1.333 -0.537 -0.393 -0.61   0.114  0.837  1.922]
```

Outliers

Un valor atípico ***outlier***, es una **observación que se desvía** significativamente de la mayoría de los datos en un conjunto.

Pueden ser el resultado de **errores** en la recolección de datos o representar **eventos inusuales** pero reales.

Es importante detectar y gestionar adecuadamente los valores atípicos:

- Pueden **distorsionar los resultados** estadísticos. Por ejemplo ciertas medidas como la media son sensibles a los valores atípicos
- **Afectar el rendimiento** de los modelos. Un valor atípico puede sesgar el modelo, haciendo que su rendimiento de predicción sea menos preciso.

Se pueden eliminar, transformar o reemplazar.

Outliers

Tipos y posibles acciones:

- **Error** en la recolección de datos, errores de entrada e incluso de codificación
  - →Se tratan en el filtrado de datos, se eliminan o se marcan como datos faltantes
- Datos en un **acontecimiento extraordinario**
  - →Se puede considerar no representativo y se podría eliminar
- Datos **extraordinarios** para los que **no se tiene explicación**
  - →Hacer el experimento con y sin estas observaciones y analizar la influencia sobre los resultados.
- **Conjunto** de observaciones cuyos **valores se desvía de la norma** formando un patrón distinto aunque cada valor podría parecer “normal”
  - →Se podría tener en cuenta pero es importante analizar cómo influyen.

Outliers

Problemática

1. Son elementos que pueden no ser representativos de la población pudiendo distorsionar seriamente el comportamiento de los contrastes estadísticos.

2. Pueden ser indicativos de las características de un segmento de la población no tenido en cuenta, por lo que puede representar falta de representatividad de la muestra.

Outliers

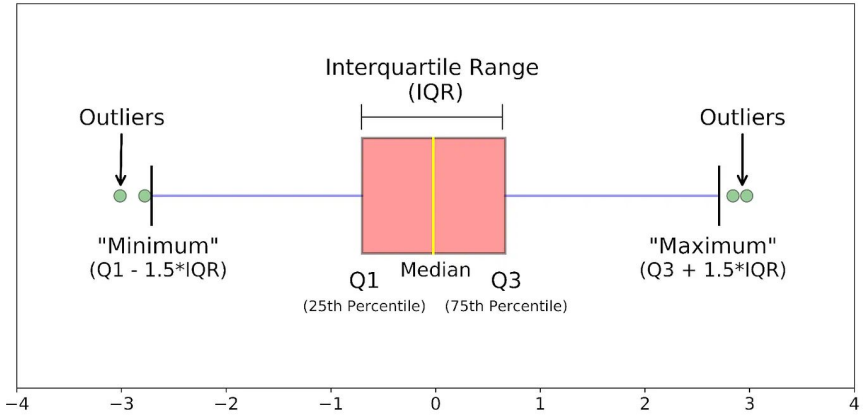
Gestionar los valores atípicos

```
edad = [5, 16, 9, 18, 15, 25, 35, 50, 52, 45, 48];

edadMedia = 28.9


edad = [5, 16, 9, 18, 15, 25, 35, 50, 52, 45, 48, 300];

edadMedia = 42.25

//La media quizá ya no es un valor representativo del conjunto
```

Outliers

Gestionar los valores atípicos

Diagrama de cajas y rango intercuartílico (IQR)



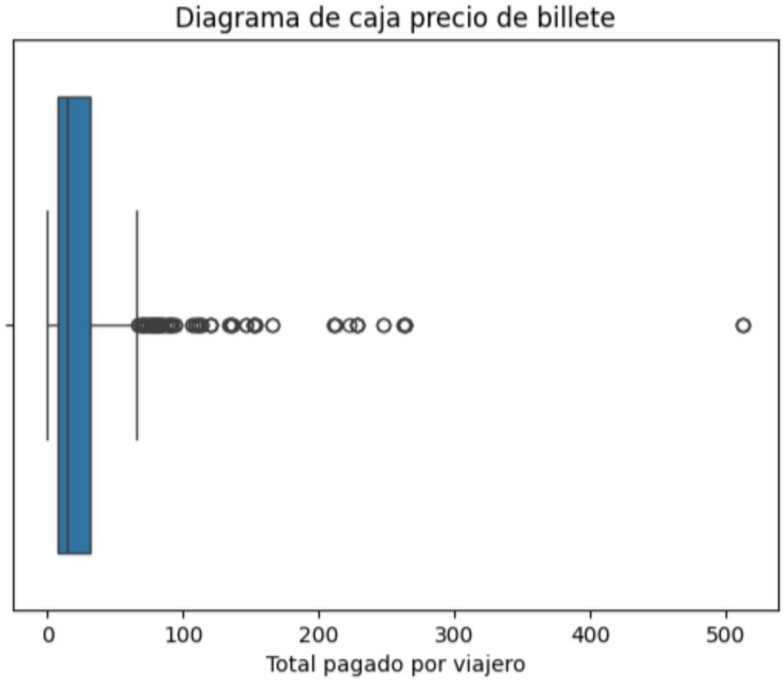
Q1 primer cuartil 25% Q3 tercer cuartil 75% IQR = Q3-Q1 Límite inferior = Q1-1,5*IQR Límite superior = Q3+1,5*IQR	Outliers están fuera de los límites inferior o superior

Outliers

Gestionar los valores atípicos *Outliers*

Diagrama de cajas

```
# Crear un diagrama de caja
g = sns.boxplot(data = titanic, x = 'fare')

# Título y eje X
g.set_title('Diagrama de caja precio de billete')
g.set_xlabel('Total pagado por viajero')
```



Outliers

Gestionar los valores atípicos *Outliers* - Rango intercuartílico (IQR)

```
# Rango intercuartílico (IQR)
# Calcular Q1 Q3 percentiles
q3 = titanic['fare'].quantile(0.75)
q1 = titanic['fare'].quantile(0.25)

# Obtener IQR
iqr = q3 - q1

# Limites superior e inferior
limitSup = q3 + (1.5 * iqr)
limitInf = q1 - (1.5 * iqr)

# Obtener el subconjunto de este dataset
outliers = titanic[(titanic['fare'] < limitInf) | (titanic['fare'] > limitSup)]
outliers.head()

#Obtener número de outliers
print(f'Outliers: {len(outliers)}')
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
27	0	1	male	19.0	3	2	263.0000	S	First	man	True	C	Southampton	no	False
31	1	1	female	NaN	1	0	146.5208	C	First	woman	False	B	Cherbourg	yes	False
34	0	1	male	28.0	1	0	82.1708	C	First	man	True	NaN	Cherbourg	no	False
52	1	1	female	49.0	1	0	76.7292	C	First	woman	False	D	Cherbourg	yes	False

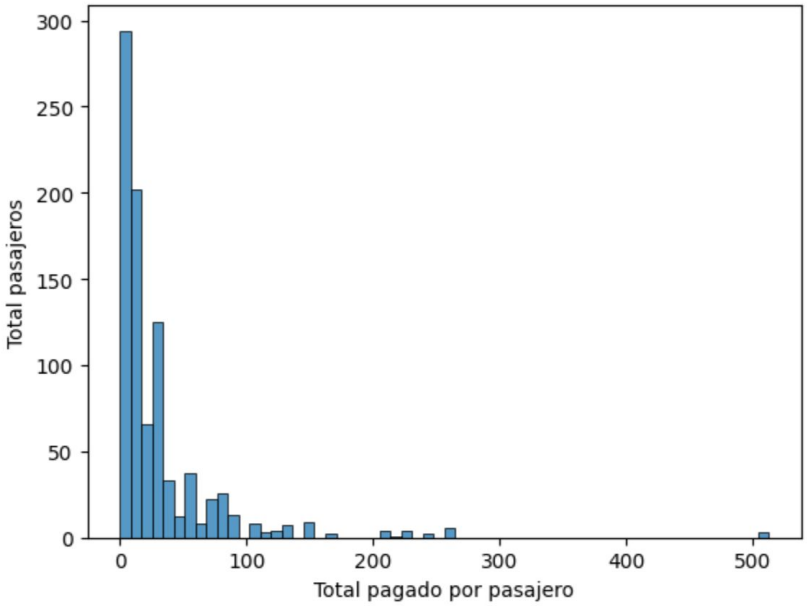
```
Outliers: 116
```

Outliers

Gestionar los valores atípicos *Outliers*

Histograma

```
# Crear histograma
g = sns.histplot(data = titanic, x = 'fare')

# Añadir etiquetas
g.set_xlabel('Total pagado por pasajero')
```



Estrategias y técnicas de limpieza, transformación y normalización

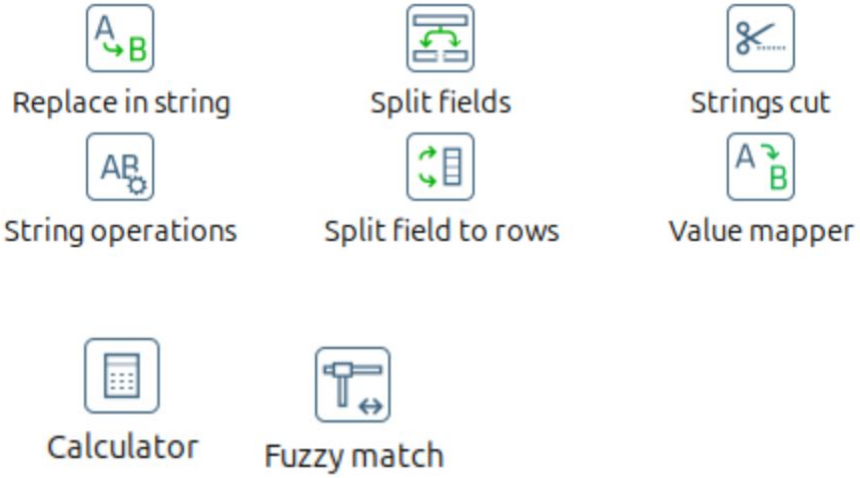
Estrategias y técnicas para la limpieza de datos

- Uso de expresiones regulares para encontrar y reemplazar formatos específicos.
- Limpieza para eliminar espacios en blanco, caracteres especiales, puntuación no deseada o caracteres extraños.
- Algoritmos para la comparación de cadenas (*string matching*), se identifican errores tipográficos, duplicados.
- Validación con datos del dominio →Mapeo de datos a vocabularios controlados.
- Estandarización aplicando un formato único o valor de referencia.

PDI Limpieza, transformación y normalización

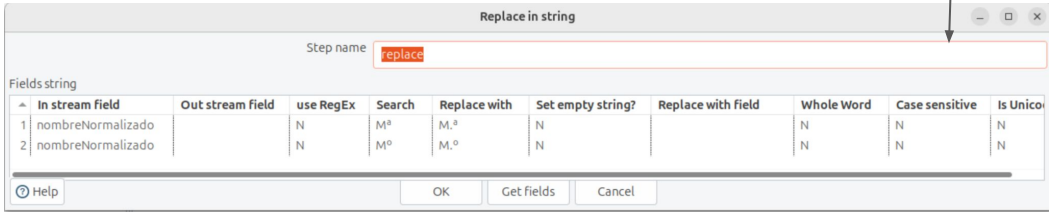
Algunos de los pasos más frecuentes para la limpieza de datos:

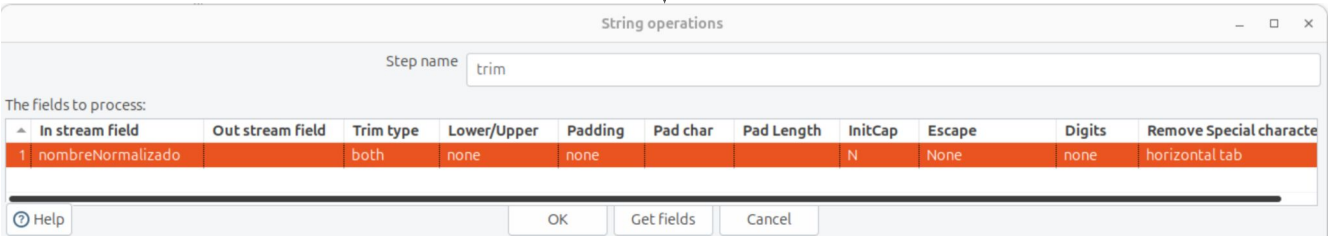
- *Replace in string*
- *Split fields*
- *Split field to rows*
- *Strings cut*
- *String operations*
- *Value mapper*
- *Calculator*
- *Fuzzy match*



PDI Limpieza, transformación y normalización

Ejemplo limpieza, transformación y normalización de datos

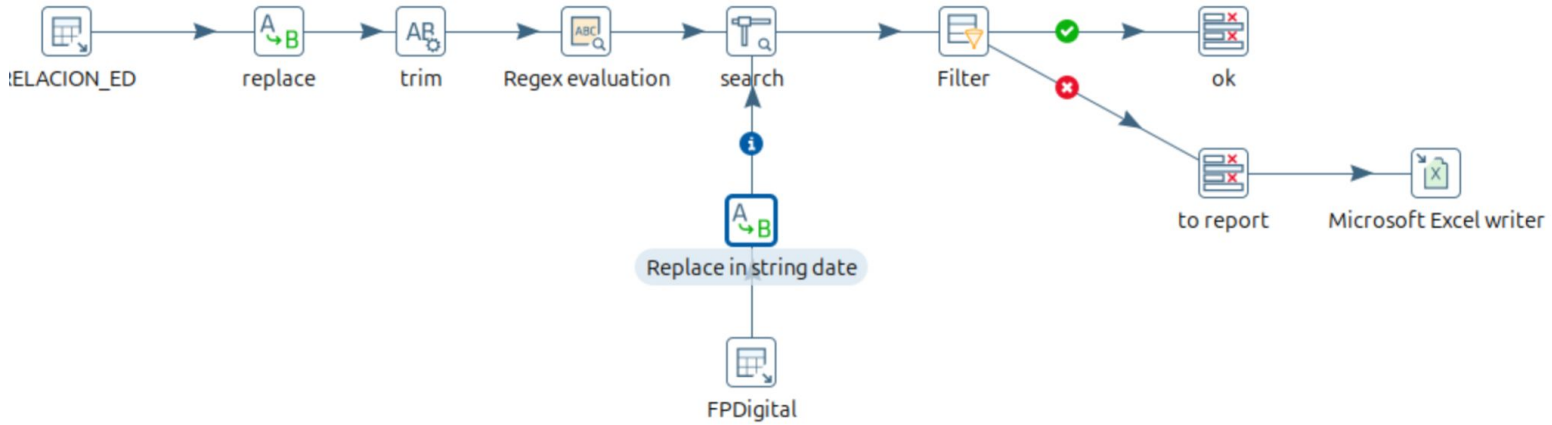




PDI Limpieza, transformación y normalización

Ejemplo limpieza de datos

Limpia el stament LPO de un registros y elimina datos que no deben estar en él
como las páginas pp. los ()
Después extrae el año de publicación original, en este caso no se desea fechas como 1985-2-8 sino solo 1985



String Matching Algorithms

String Matching Algorithms

1. **Levenshtein y Damerau-Levenshtein**: Calcula la distancia entre dos strings en función del número de cambios necesarios para pasar de una cadena a otra.

2. **Needleman-Wunsch:** Calcula la similitud de dos secuencias, calculando una penalización por espacio.

3. **Jaroa y Jaro-Winkler**: Calcula un índice de similitud entre dos strings. El resultado es una fracción entre 0 (sin similitud) y 1 (coincidencia idéntica).

4. **Pair letters similarity**: Este algoritmo corta ambas cadenas en pares y comparan los conjuntos de pares.

5. **Metaphone, Double Metaphone, Soundex y RefinedSoundEx**: Estos algoritmos todos intentan hacer coincidir los strings en función de cómo "sonarían" y también se denominan fonéticas algoritmos.

String Matching Algorithms

String Matching Algorithms

Levenshtein y Damerau-Levenshtein: Calcula la distancia entre dos strings en función del número de cambios necesarios para pasar de una cadena a otra.

- CASTERS y CASTRO es de 2
- Paso 1: eliminar la E; Paso 2: reemplazar S por O

Needleman-Wunsch: Calcula la similitud de dos secuencias, calculando una penalización por espacio.

- Puntuación de -2 para CASTERS a CASTRO

String Matching Algorithms

String Matching Algorithms

Jaroa y JaroWinkler: Calcula un índice de similitud entre dos strings. El resultado es una fracción entre 0 (sin similitud) y 1 (coincidencia idéntica).

- Levenshtein entre CASTERS y POOH es de 7
- Jaro y de Jaro-Winkler es 0 porque no hay similitud entre las dos cadenas

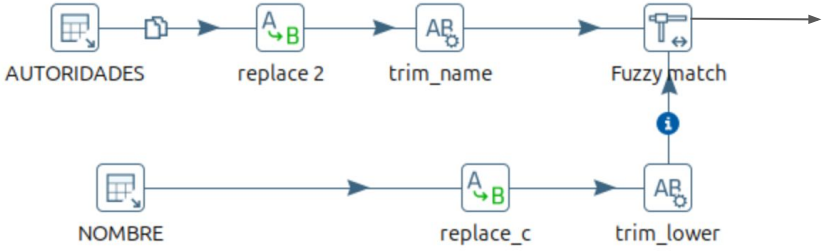
Pair letters similarity: Este algoritmo corta ambas cadenas en pares y comparan los conjuntos.

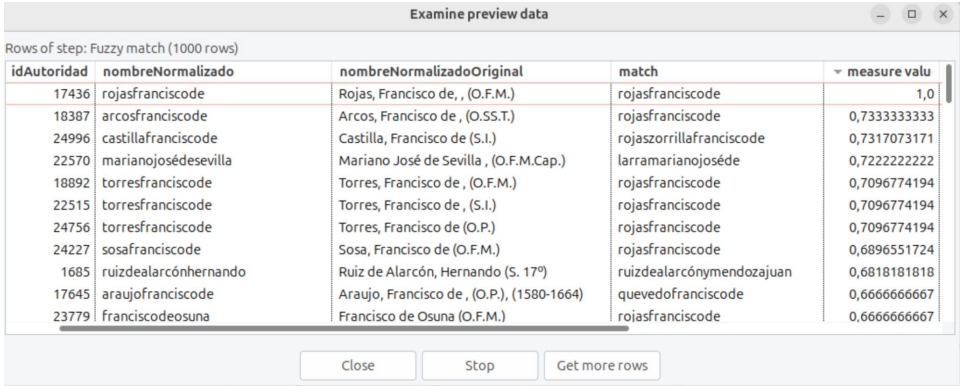
- CASTERS y CASTRO se transformará en:
  - {CA, AS, ST, TE, ER, RS} y {CA, AS, ST, TR, RO}
  - Calcula la similitud dividiendo el número de pares comunes (multiplicado por dos) por la suma de los pares de ambas cadenas → hay tres pares comunes (CA, AS, ST) y once pares en total →puntuación de similitud de (2*3)/11 = 0,545 (está bien)

String Matching Algorithms

Objetivo: determinar si dos nombres de autor, que parecen ligeramente diferentes, en realidad se refieren a la misma persona

Si la distancia es baja →asumimos que son duplicados





Examine preview data

Rows of step: Fuzzy match (1000 rows)

idAutoridad	nombreNormalizado	nombreNormalizadoOriginal	match	measure valu
17436	rojasfranciscode	Rojas, Francisco de, , (O.F.M.)	rojasfranciscode	1,0
18387	arcosfranciscode	Arcos, Francisco de , (O.SS.T.)	rojasfranciscode	0,7333333333
24996	castillafranciscode	Castilla, Francisco de (S.I.)	rojaszorrillafranciscode	0,7317073171
22570	marianojosédesevilla	Mariano José de Sevilla , (O.F.M.Cap.)	larramarianojoséde	0,7222222222
18892	torresfranciscode	Torres, Francisco de , (O.F.M.)	rojasfranciscode	0,7096774194
22515	torresfranciscode	Torres, Francisco de , (S.I.)	rojasfranciscode	0,7096774194
24756	torresfranciscode	Torres, Francisco de (O.P.)	rojasfranciscode	0,7096774194
24227	sosafranciscode	Sosa, Francisco de (O.F.M.)	rojasfranciscode	0,6896551724
1685	ruizdealarcónhernando	Ruiz de Alarcón, Hernando (S. 17º)	ruizdealarcónymendozajuan	0,6818181818
17645	araujofranciscode	Araujo, Francisco de , (O.P.), (1580-1664)	quevedofranciscode	0,6666666667
23779	franciscodeosuna	Francisco de Osuna (O.F.M.)	rojasfranciscode	0,6666666667

Close Stop Get more rows

String Matching Algorithms

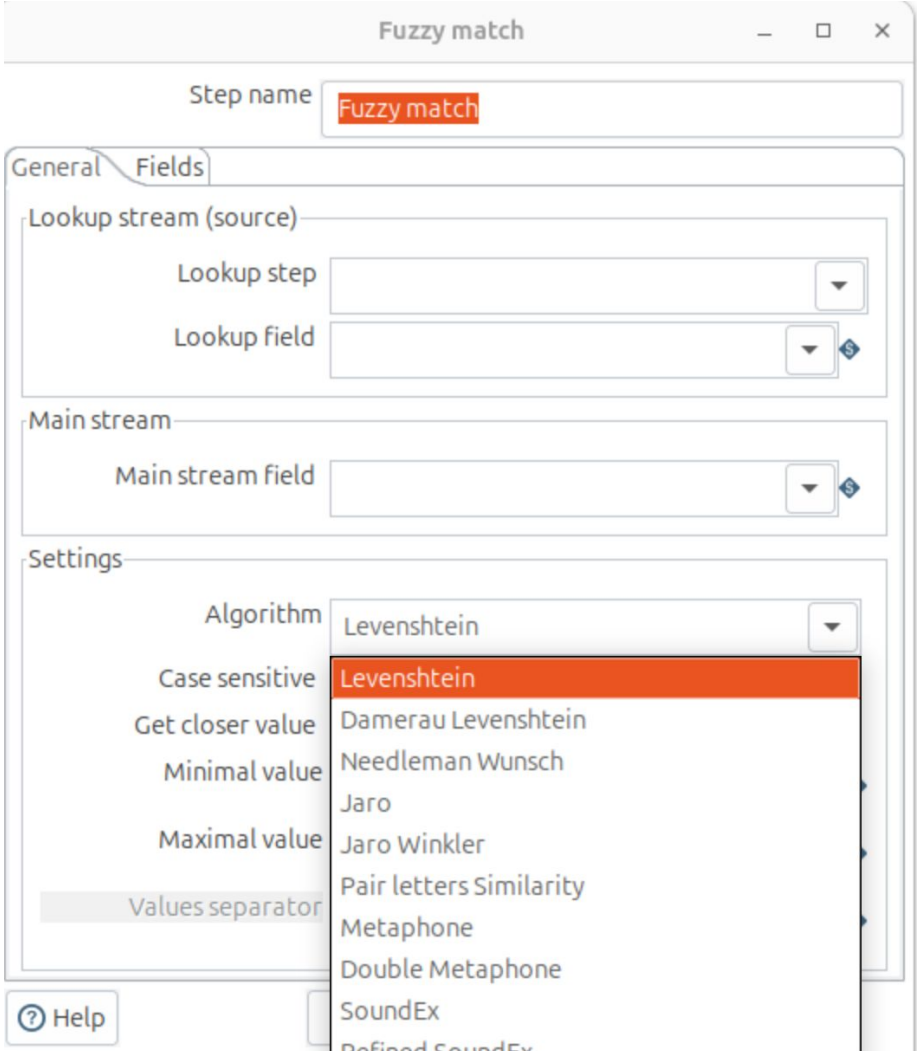
String Matching Algorithms

Metaphone, Double Metaphone, Soundex y RefinedSoundEx: Estos algoritmos intentan hacer coincidir los strings en función de cómo "sonarían", también se denominan algoritmos fonéticos.

La debilidad de todos estos algoritmos fonéticos es que se basan en el inglés y no sería de mucha utilidad en un entorno francés, español u holandés.

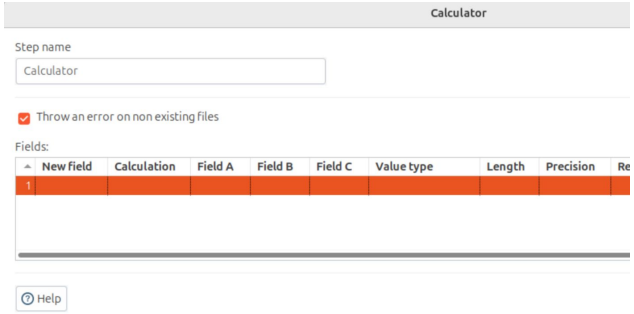
String Matching Algorithms

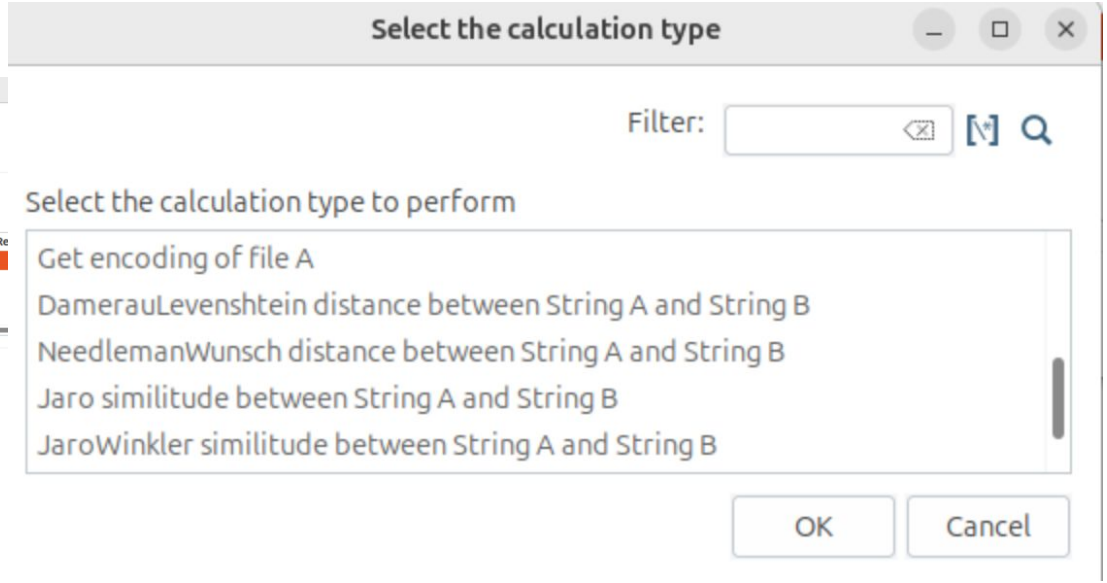
Fuzzy match



String Matching Algorithms

Calculator step





Estrategias y técnicas para la limpieza de datos

- Uso de expresiones regulares para encontrar y reemplazar formatos específicos.
- Limpieza para eliminar espacios en blanco, caracteres especiales, puntuación no deseada o caracteres extraños.
- Algoritmos para la comparación de cadenas (*string matching*), se identifican errores tipográficos, duplicados.
- Validación con datos del dominio →Mapeo de datos a vocabularios controlados.
- Estandarización aplicando un formato único o valor de referencia.

Validación de datos

Objetivo: verificar y validar los datos recopilados, garantizando que son **confiables** y **adecuados**.

Deben cumplir las **reglas** de negocio predefinidas y marcar o rechazar cualquier registro que no cumpla estos requisitos.

Cualquier tarea en la gestión de datos (adquisición, transformación o carga), debe incluir la validación de datos para garantizar resultados precisos:

- verificar que los datos se extraigan correctamente de las fuentes de origen
- la transformación se realice con precisión de acuerdo con las reglas definidas
- y la carga de datos en destino se produzca sin pérdida ni corrupción

Una forma sencilla de validación de datos son las tablas de referencia, donde la regla es: “el valor debe estar presente en tabla de referencia”.

La mayoría de las reglas de validación son un poco más complejas y tienen dependencia del dominio.

Validación de datos

- Verificar si un dato debe estar dentro de un **rango de valores**
- Verificar **consistencia** en el conjunto de datos y entre diferentes conjuntos de datos
- Verificar **Integridad referencial** garantizando que se mantengan las relaciones entre los datos de diferentes tablas o bases de datos
- Verificar las **restricciones de identidad**, garantizando que los datos como identificadores, DNI, email o teléfono, son únicos
- Comprobar que los **datos esenciales** están **presentes**
    - gestión de valores faltantes vista anteriormente

Validación de datos

- Comprobar **tipo** de datos y **tamaño**
  - por ejemplo una contraseña de 12 caracteres alfanuméricos y con ciertas restricciones, un código postal, email
- Comprobar el **formato**
  - por ejemplo la fecha formato AAAA-MM-DD o números de teléfono
- Comprobar valor dentro de un **vocabulario controlado**
  - por ejemplo los idiomas tiene un conjunto de datos aceptables, los formatos de una obra tiene una lista o tabla de referencia que determina los valores válidos, los meses del año, el género, latitud, longitud, etc.
- **Restricciones** según **dominio**
  - Los valores de entrada deben estar en mayúsculas/minúsculas
  - Los importes no pueden superar el valor X
  - Los suscriptores deben tener al menos 18 años

Validez de los datos

Tablas de referencia

Using Reference Tables

Comprobar información consultando fuentes externas validadas (datos maestros o de referencia).

No siempre lo tenemos disponible.

Por ejemplo países, ciudades, autores, libros, idiomas, etc.

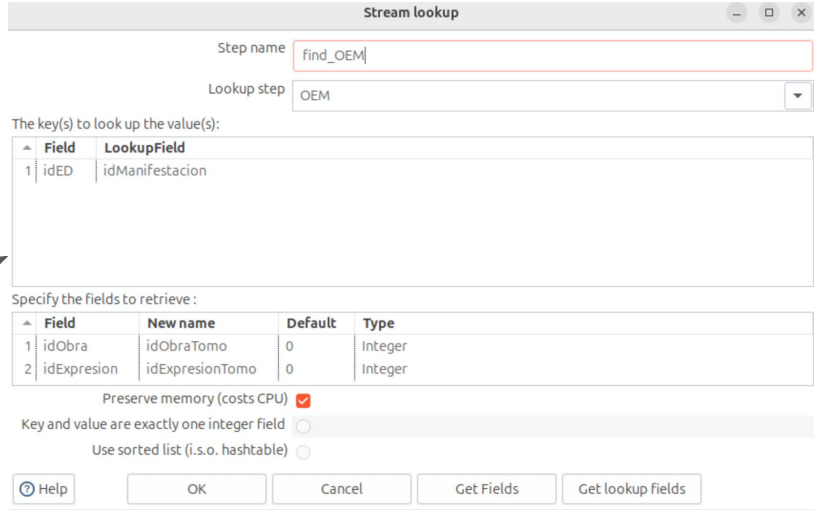
Wikidata

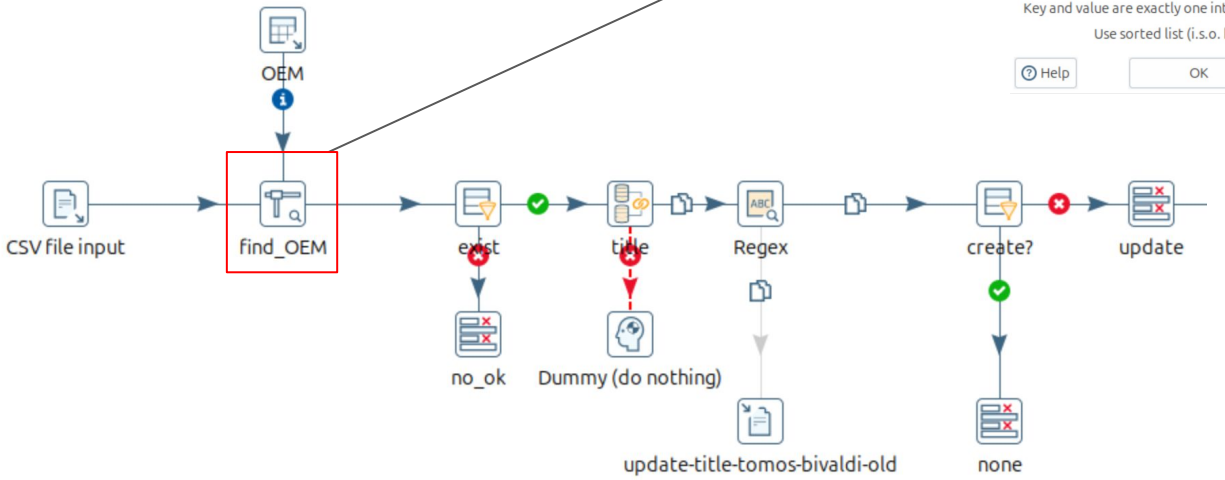
Geonames

Tablas de referencia

Conforming Data Using Lookup Tables

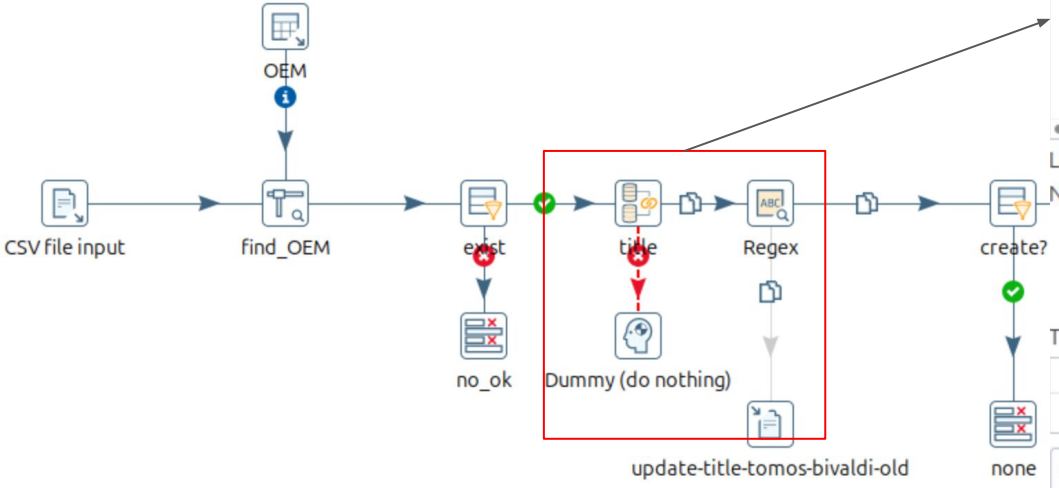
Realizar una búsqueda basada en coincidencia de campos





Tablas de referencia

Conforming Data Using Lookup Tables



Database join

Step name: title

Connection: LOCAL | Edit... | New... | Wizard...

SQL

```
SELECT
  t.idTitulo
, t.titulo as titleTomoOld
FROM TITULO t
WHERE
t.FK_idEntidadDocumental = ? AND
tipoTitulo = 1
```

Line 1 Column 0

Number of rows to retur: 0

Outer join? ☐

Replace variables ☑

The parameters to use:

	Parameter fieldname	Parameter Type
1	idObraTomo	Integer

Help | OK | Cancel | Get Fields

Tablas de referencia

Se crea tabla de referencia como tabla maestra de comprobación de datos.

Se deben cumplir dos requisitos básicos:

- Cada posible valor del sistema fuente necesita una correspondencia
- La correspondencia debe conducir a un conjunto único de valores

Ejemplo: se crea tabla de referencia para los idiomas con las normas ISO sobre identificación del idioma

- ISO 639-2 estándar internacional que define los código de identificación única de los idiomas
- ISO 639-3 define código para para un idioma individual incluyendo idiomas vivos, extintos y antiguos
- Norma ISO 639-1 es más restrictiva que las otras normas que cubren una gama más amplia de idiomas y variaciones

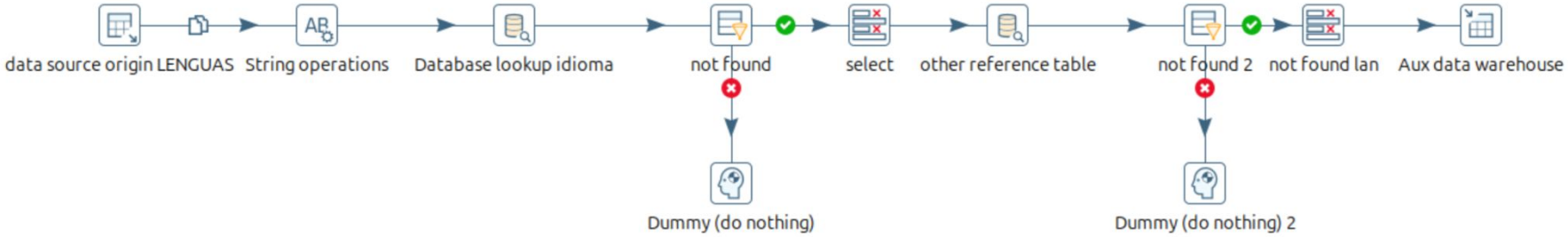
idIdioma ▲ 1	iso6392Code	iso6393Code	iso6391Code	nombre
26	spa	spa	es	español
27	swe	swe	sw	sueco
28	syr	syr		sirio
30	lat	lat	la	llatí
32	gre* / ell	ell	el	grec
33	gre* / ell	ell	el	griego
37	syc	syc		siriaco
38	spa	spa	es	castellà
39	cat	cat	ca	català
45	pol	pol	pl	polaco
46	cat	cat	ca	valenciano
47	baq* / eus	eus	eu	euskera
48	dan	dan	da	danés

Tabla de referencia del código de idioma

Tablas de referencia

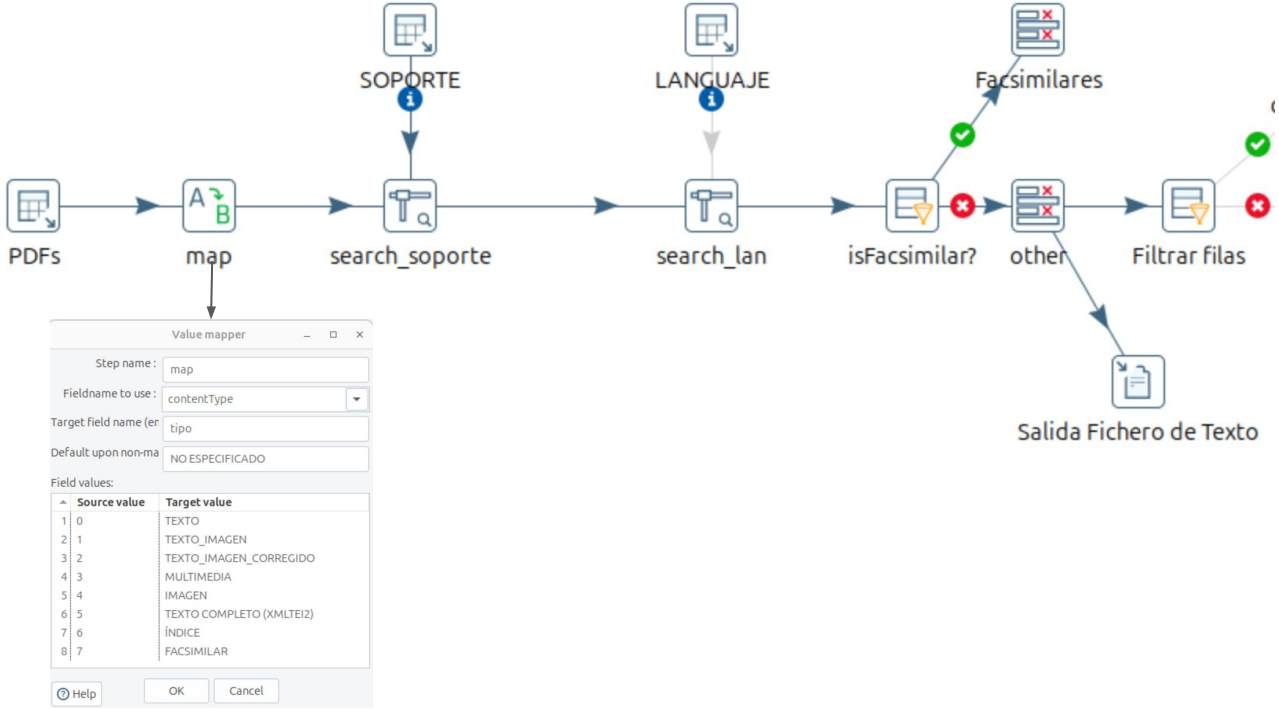
Conforming Data Using Reference Tables

Ejemplo



Validación de datos

Mapeo de datos a vocabularios controlados



Validación de datos

Applying Validation Rules

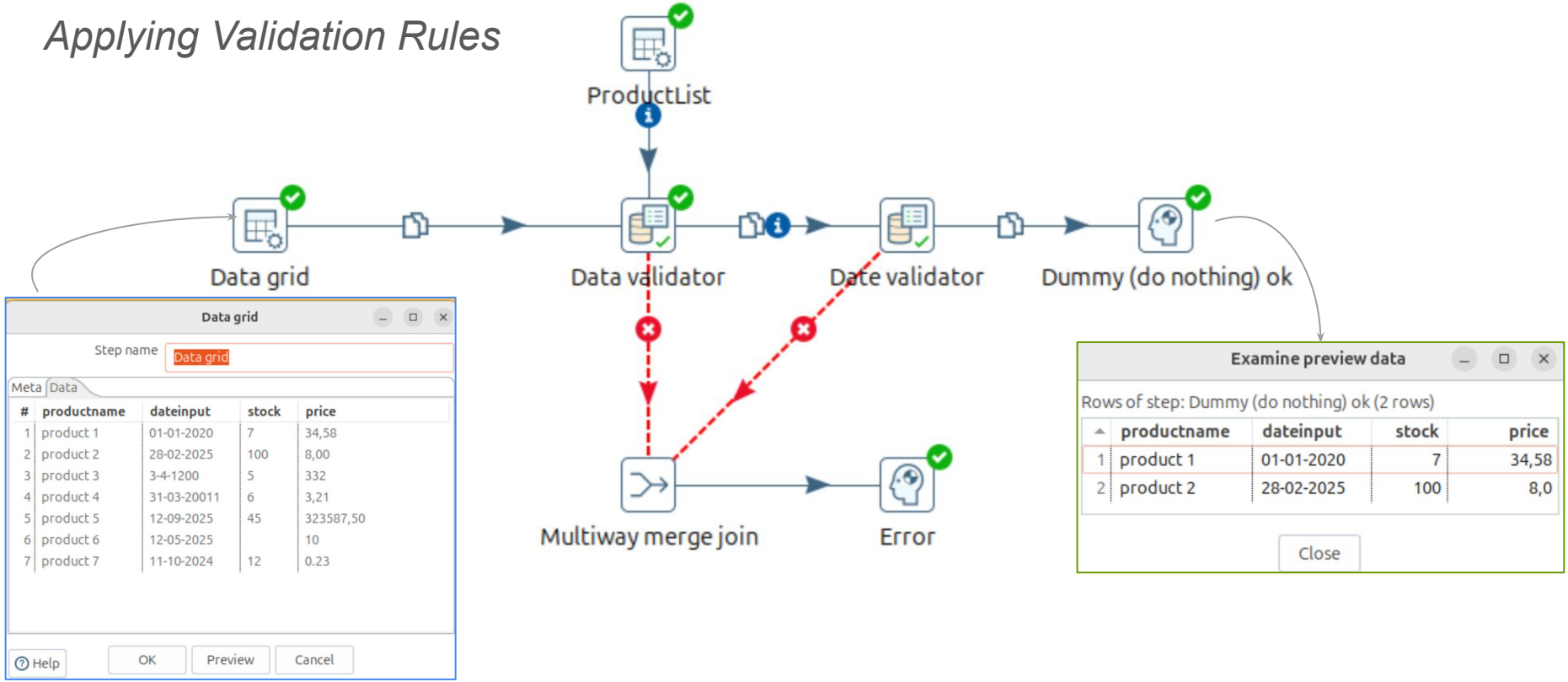
Ejemplo:

Los datos deben cumplir las siguientes reglas:

- Ninguno de los campos puede contener un valor NULL.
- Las fechas no pueden ser anteriores al 1 de enero de 2001.
- No se permiten nombres de productos fuera del catálogo que define la empresa.
- El número de artículos en stock debe estar entre 1 y 100. No puede estar vacío.
- El precio por artículo no puede exceder de 1000€.

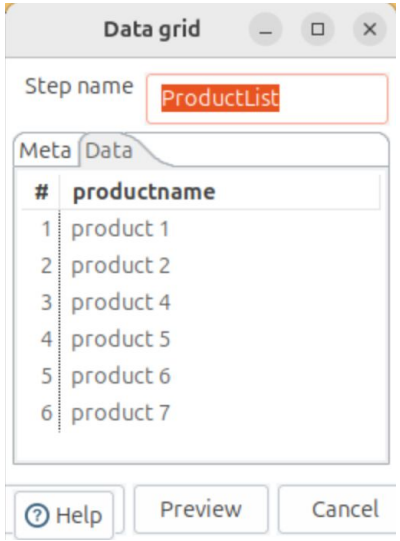
Validación de datos

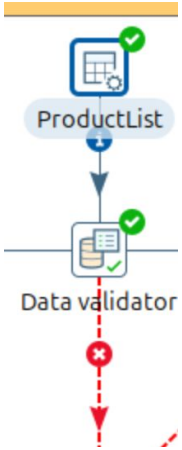
Applying Validation Rules

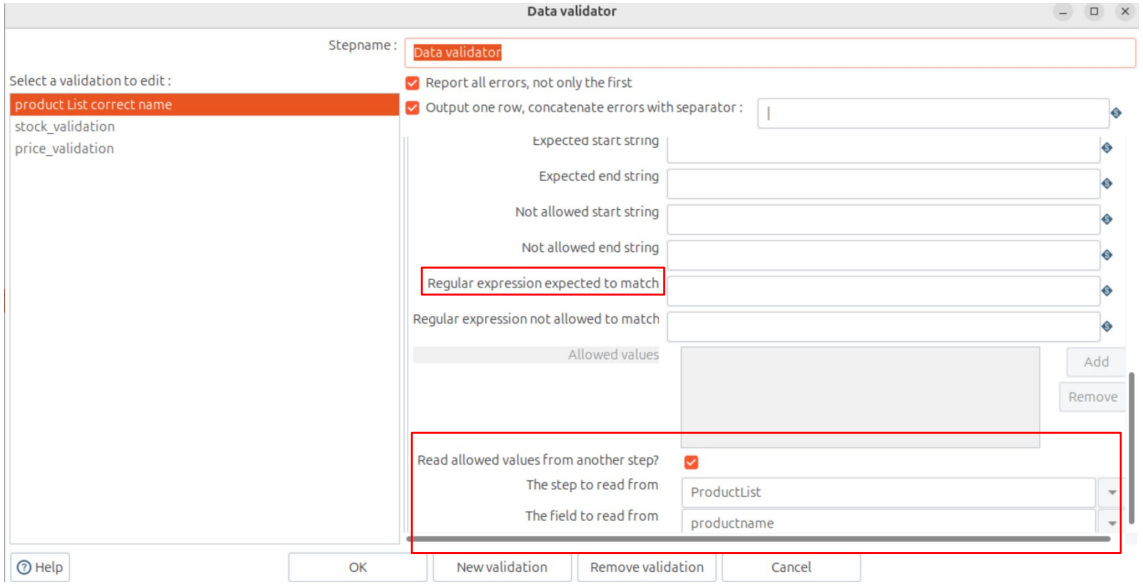


Validación de datos

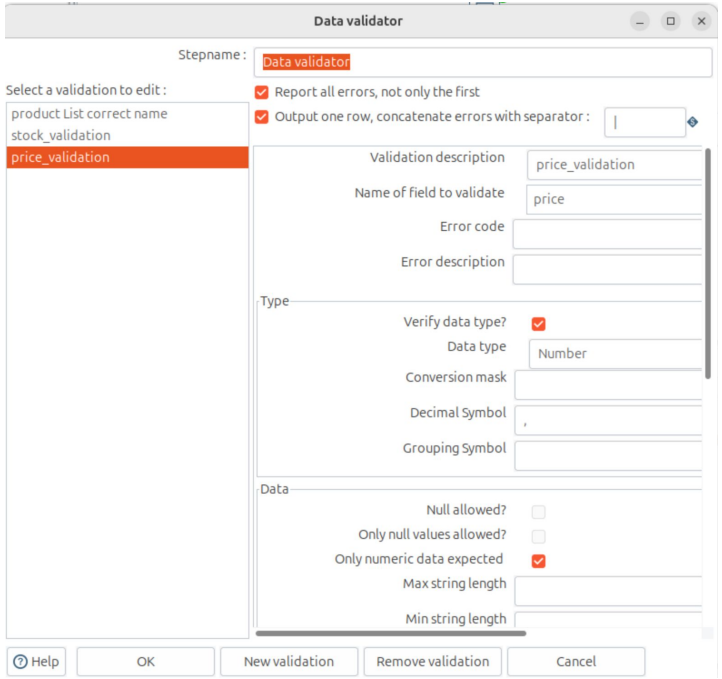
Applying Validation Rules

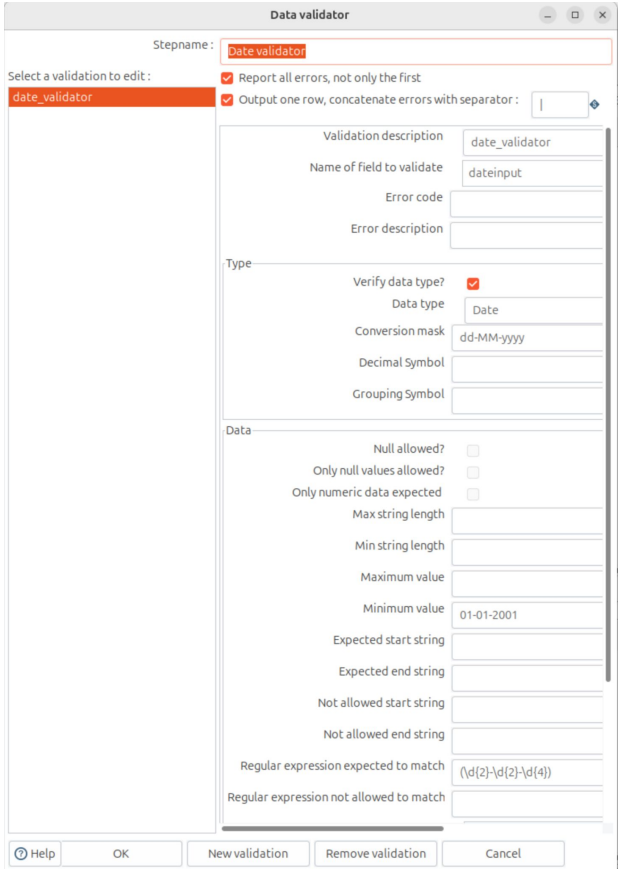






Validación de datos





Gestión de datos duplicados

El tratamiento de duplicados se basa en **identificar y eliminar esa redundancia** en base a los **atributos clave** que definen qué conforma un duplicado y las **reglas** de negocio propias del dominio

- Diversas fuentes de datos: Obtener la misma información de diferentes sistemas que no están sincronizados.
- Errores en la entrada de datos: Errores tipográficos que crean nuevas entradas para la misma entidad.
- Fusión de bases de datos: Combinar conjuntos de datos sin un proceso adecuado de deduplicación.

Problemas derivados:

- Baja calidad de los datos
- Inconsistencia ya que se pueden contener información contradictoria
- Espacio de almacenamiento no necesario
- Impacto en el rendimiento de los modelos

Gestión de datos duplicados

Resolver duplicados en el conjunto de datos se realiza durante la fase de transformación del proceso ETL.

El objetivo es **identificar y eliminar** los registros **duplicados** para mantener solo la versión más precisa y completa.

Pasos clave:

1. **Identificar**: encontrar los registros que son potencialmente duplicados.
    - Duplicados exactos
    - Duplicados no exactos
2. **Estandarizar**: **normalizar** los datos para facilitar la comparación
3. **Fusionar** (*merge*) o **eliminar** (*delete*)

Gestión de datos duplicados

Identificar

Encontrar los registros que son potencialmente duplicados.

- Duplicados exactos, usando identificadores únicos
- Duplicados no exactos, con técnicas *fuzzy matching* para encontrar coincidencias aproximadas. Ejemplos de uso de Jaro-Winkler o Levenshtein para calcular la similitud entre cadenas de texto

Herramientas:

- Herramientas ETL comerciales
- Librerías de código abierto: Python (Pandas, Dask, Dedupe)
- SQL usando por ejemplo GROUP BY

Gestión de datos duplicados

Estandarizar

Normalizar los datos para facilitar la comparación

Mayúsculas/minúsculas

Limpieza de datos

ej. "calle Mayor", "C/ Mayor", "C. Mayor" se convierten en "Calle Mayor"

Gestión de datos duplicados

Fusionar o eliminar

Una vez identificados, decidir qué hacer con los duplicados:

- **Fusionar** (Merge): Combinar la información de los registros duplicados en un único registro maestro, conservando los datos más completos y actualizados.
- **Eliminar** (Delete): Si no se puede fusionar, se elimina el duplicado menos relevante.

Procesos ETL

Tema 7: Diseño de flujos de trabajo a través de jobs

Adquisición y preparación de datos

IA EPS UA

Universitat d'Alacant
Universidad de Alicante

Tema 7: Diseño de flujos de trabajo a través de jobs

- Diseño de jobs
- Variables y parámetros

Ciclo de vida del proyecto IA

Flujos de trabajo

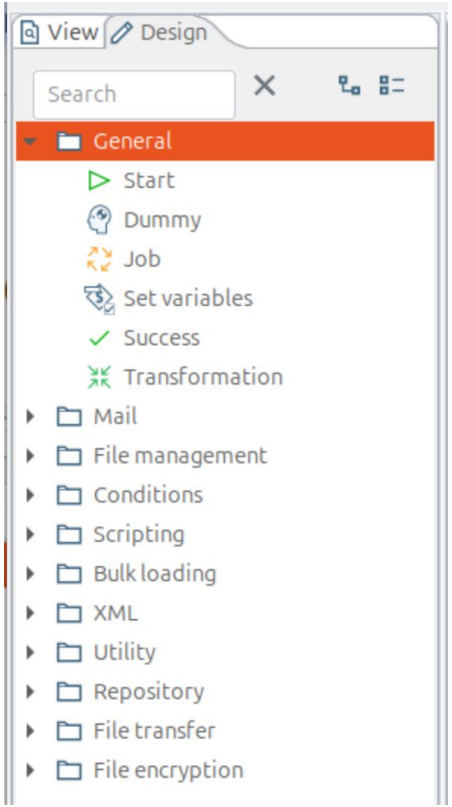
Existen diferentes herramientas que permiten trabajar con grandes volúmenes de datos y facilitan su análisis.

Para ello se diseñan y desarrollan flujos de trabajo que se pueden crear utilizando pasos o entradas unidas por saltos que pasan datos de un elemento al siguiente.

Los **flujos de trabajo** utilizan dos tipos de archivos:

1. Transformaciones que realizan tareas ETL.
2. Trabajos que organizan actividades de ETL, como la definición del flujo, dependencias y ejecución.

Pentaho Data Integration



Pasos para el diseño de *jobs*

General, pasos de inicio de trabajo, ejecución transformaciones o trabajos entre otras operaciones

Mail, envío de correos, recuperación y validación de cuentas

File management, gestión de operaciones sobre ficheros y carpetas, HTTP, etc

Conditions, comprobaciones sobre conexiones, BD, ficheros, etc

Scripting, scripts en JavaScript, Shell y SQL

Bulk loading, cargas a MySQL y MSSQL

XML, validaciones XML, DTD, XSD y XSL

Utility, pasos para ejecución de transformaciones como ping, enviar información a syslogs, Nagios, etc

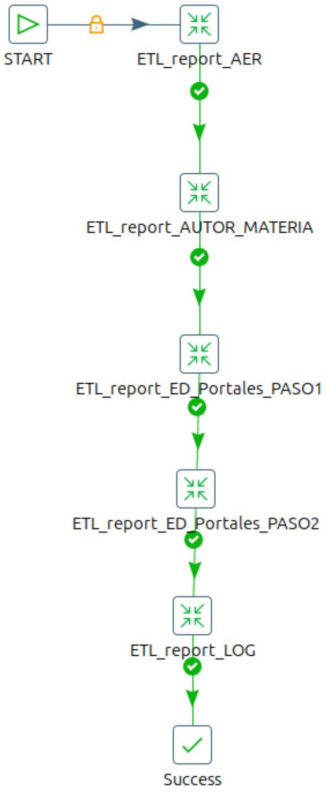
Repository, operaciones con el repositorio de transformaciones y trabajos.

File transfer, gestión ficheros FTP, FTPS y SFTP

File encryption : sistema PGP (Pretty Good Privacy) para envío y recepción de ficheros

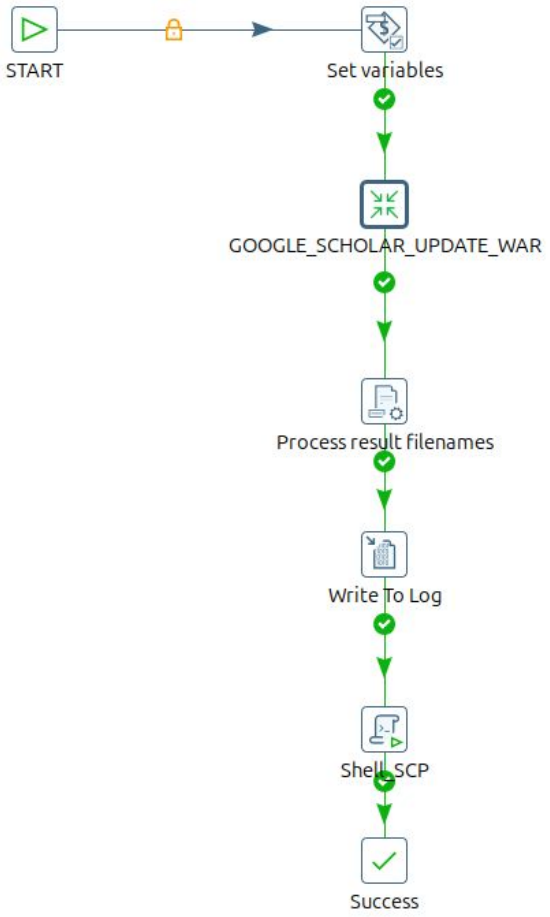
Flujos de trabajo

Job



Flujos de trabajo

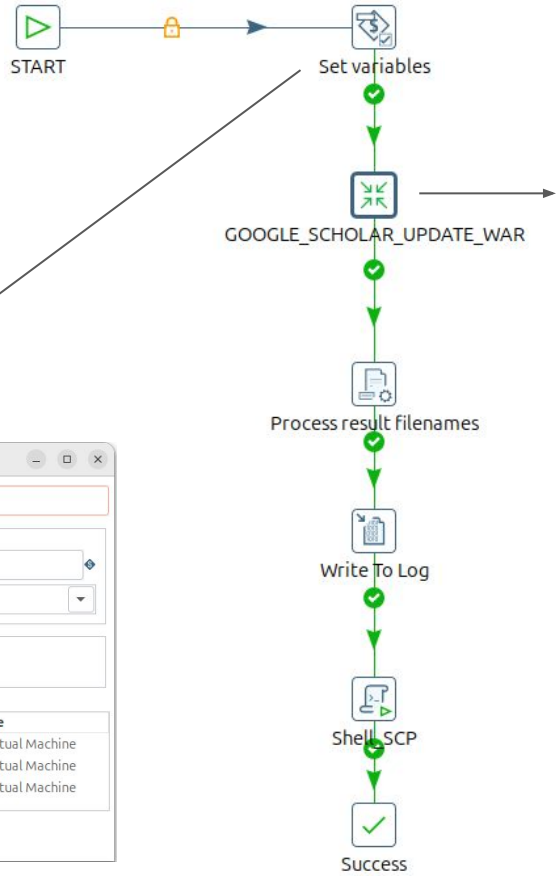
Job

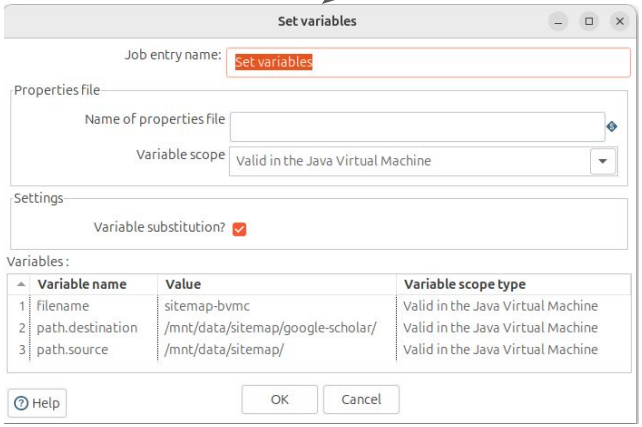


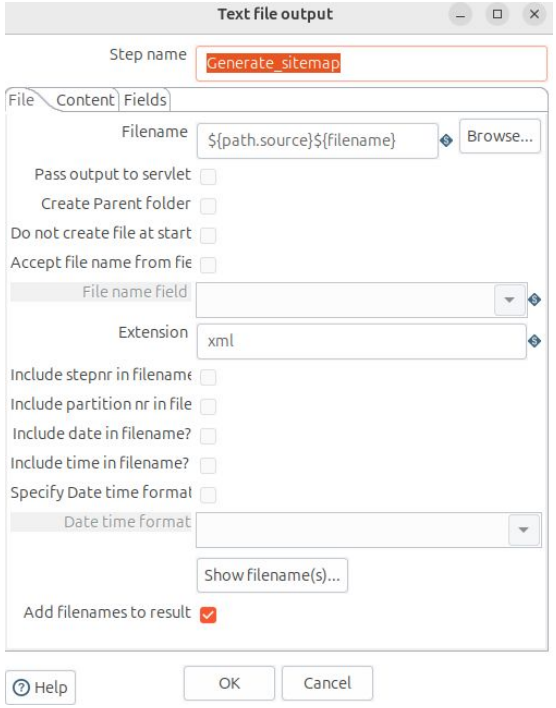
Flujos de trabajo

Job

Establecer variables del flujo de trabajo, paso

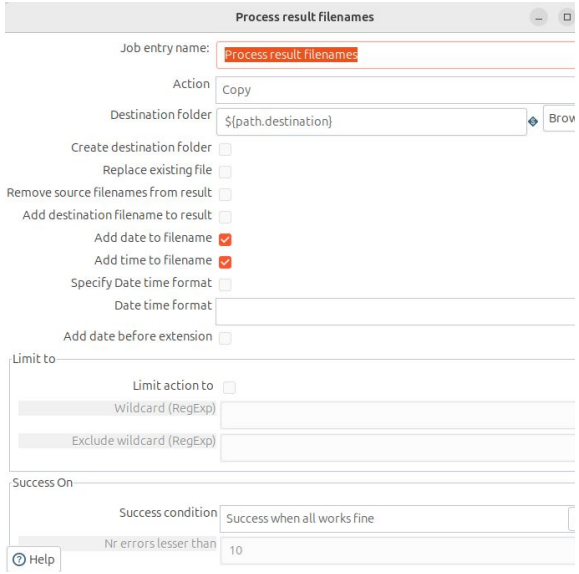


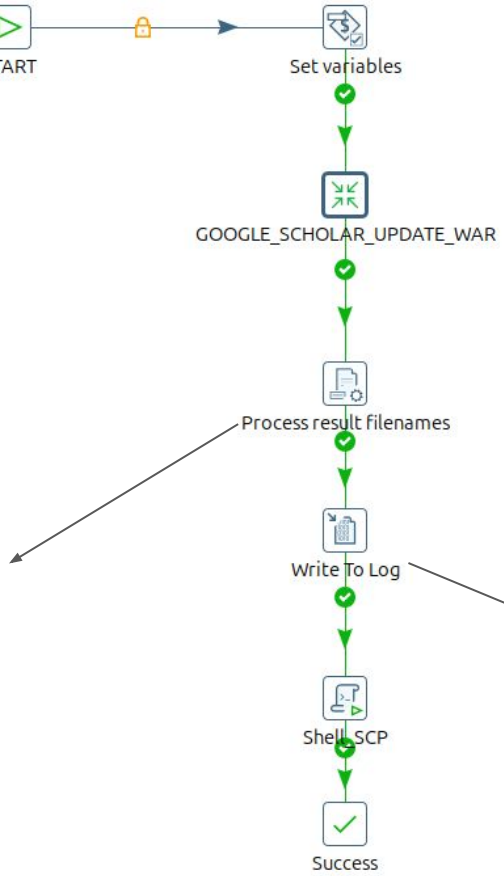


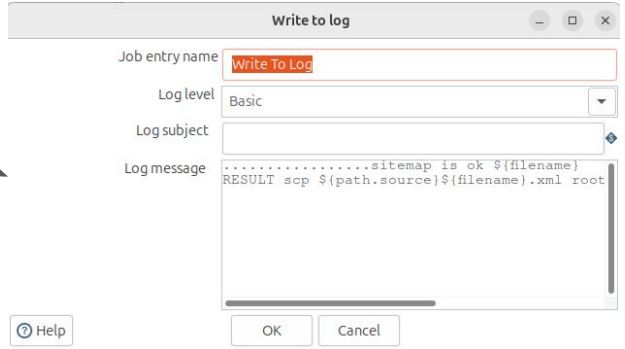


Flujos de trabajo

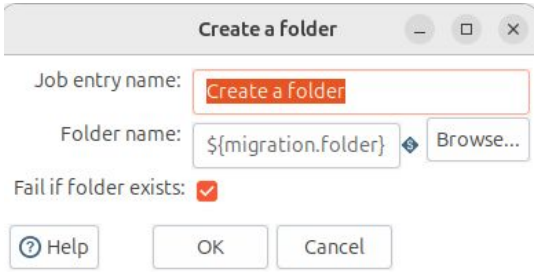
Job

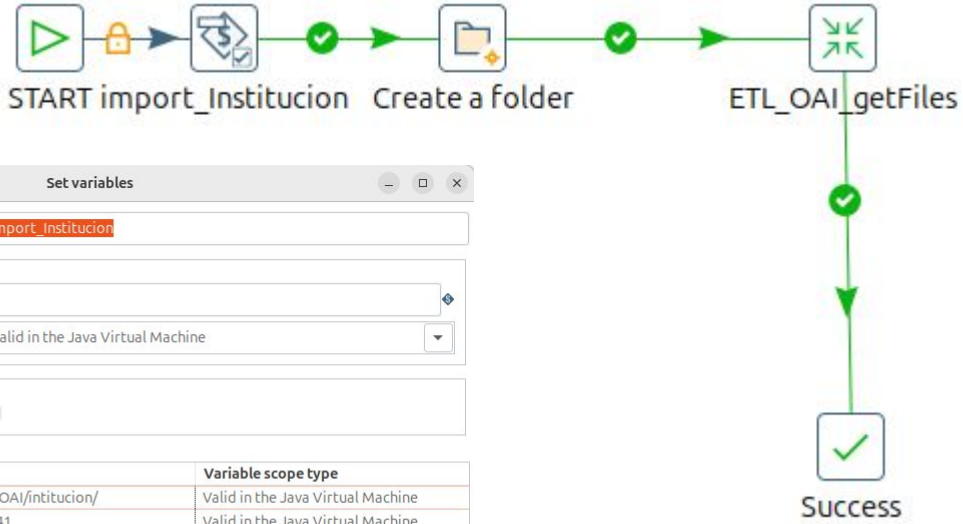


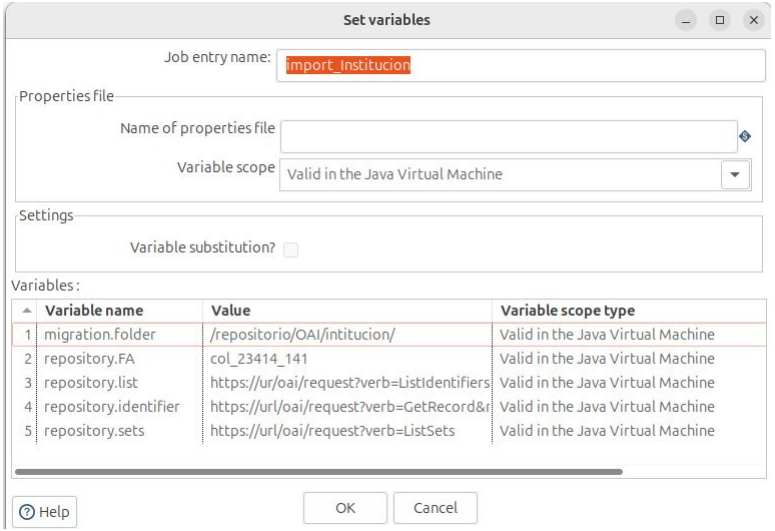




Flujos de trabajo





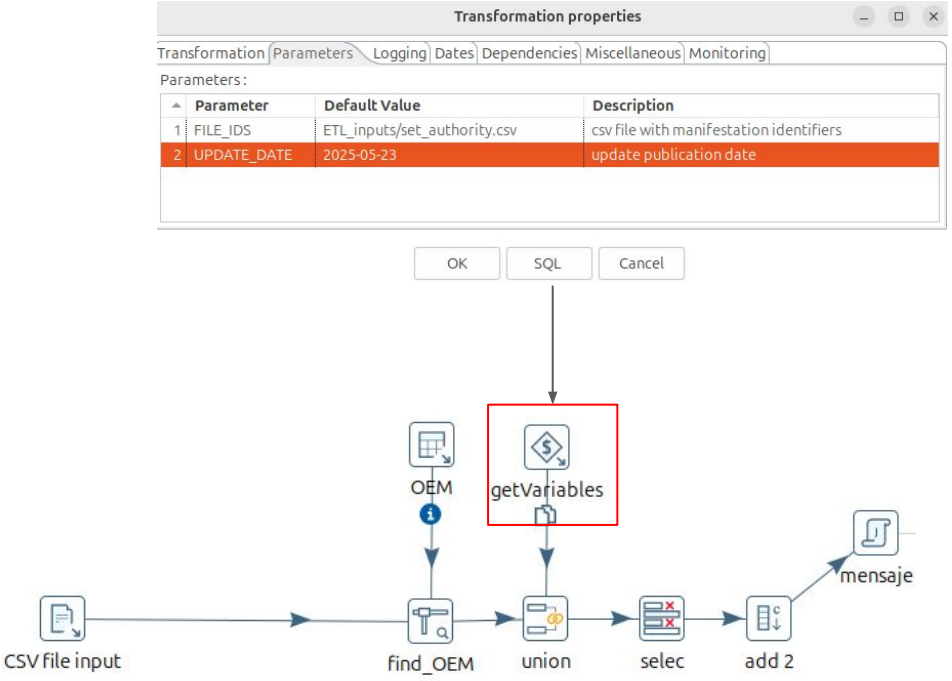


	Variable name	Value	Variable scope type
1	migration.folder	/repositorio/OAI/intitucion/	Valid in the Java Virtual Machine
2	repository.FA	col_23414_141	Valid in the Java Virtual Machine
3	repository.list	https://ur/oai/request?verb=ListIdentifiers	Valid in the Java Virtual Machine
4	repository.identifier	https://url/oai/request?verb=GetRecord&r	Valid in the Java Virtual Machine
5	repository.sets	https://url/oai/request?verb=ListSets	Valid in the Java Virtual Machine

Flujos de trabajo

Uso de variables en procesos ETL

- Las variables se configuran a nivel de sistema o dinámicamente en un trabajo
- Ámbito específico que permite ejecutar en paralelo el mismo trabajo o transformación con diferentes variables configuradas en el mismo sistema a nivel de JVM, de job actual o sesión
- Tipos de variables: sistema o definida por el usuario



Transformación y enriquecimiento de datos

Adquisición y preparación de datos

IA EPS UA

Universitat d'Alacant
Universidad de Alicante

Contenido

- Web Semántica y Linked Open Data
  - Wikidata
  - Consultas con SPARQL
- Ontologías y vocabularios/taxonomías
  - Schema.org
  - Otros ejemplos
- Transformación de datos
  - Tecnologías
  - Ejemplos
- Reutilización
  - Calidad de datos
  - Visualización
  - Retrieval-Augmented Generation

Web semántica y Linked Open Data

Slide1

Source of slides

The Semantic Web

DAML
2000 August 15

Tim Berners-Lee, Dan Connolly, Lynn Andrea Stein, Ralph Swick

http://www.w3.org/2000/Talks/0815-daml-sweb-tbl

This text corersponds approximately to the intent of the slides.

Semantic Web

The essential property of the World Wide Web is its universality. The power of a hypertext link is that "anything must be able to link to anything". This requires that anything must be able to be put on the web. The web technology, then, must not discriminate between cultures, between media, between the scribbled draft and the polished performance, between commercial and academic information, and so on. Information varies along many axes. One of these is the difference between information produced primarily for human consumption, and that produced primarily for machines. At one end of the scale we have the 5 second TV commercial and poetry. At the other end we have the database. To date, the web has developed most rapidly as a medium of documents for people rather than data which can be processes automatically. In this article we look toward the *Semantic Web*, in which data with well defined meaning is exchanged, and computers and people work side by side in cooperation.

Weblike

We are looking for a web in which data has processable qualities typical of databases and mathematical formulae, but which is still weblike. Weblike things are, like the Internet, decentralized. They typically involve a lot of fun at every level, and provide benefits on a macroscopic scale which are hard or impossible to predict in advance. To actually build such systems typically requires some compromises, just as the Web initially had to throw away the ideal of total consistency, thus ushering the infamous *Error 404 not found* message but allowing unchecked exponential uncontrolled growth. The semantic web must be weblike in that, to be universal -- to ask any information which may be useful to be easily available -- it must be minimally constraining. Like good government, it must require only that which is essential for everything to work. If you think of the web today as turning all the documents in the world into one big book, then think of the Semantic Web as turning all the data into one big database, or one big mathematical formula. The semantic web is not a separate web - it is a new form of information adding one more dimension to the diversity of the one web we have today.

First level: RDF

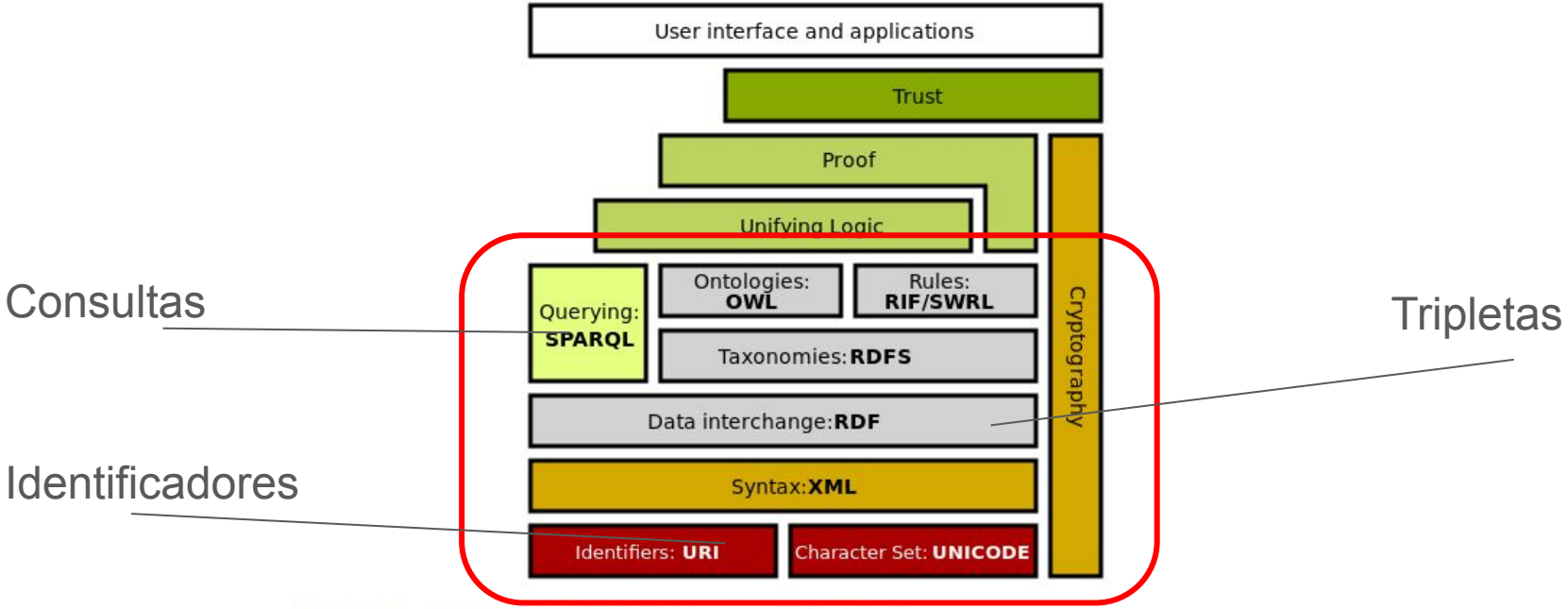
Two important technologies for the semantic web are already in place. The XML language for expressing information gives the community a common way of creating structured documents, thoug it doesn't have a way of saying in any sense what they mean. However, a starting point for talking about meaning comes from developments in the metadata world of information about information such as library cards and catalogs of web resources. This was the initial impetus behind the development of the Resource Description Framework (RDF) which provides a simple model of conveying the meaning of information as sets of triples rather like the subject, verb and object in a sentence.

Extensión de la Web tradicional para dotarla de significado y pueda ser entendida por máquinas y humanos

https://www.w3.org/2000/Talks/0906-xmlweb-tbl/text.htm

Web semántica y Linked Open Data

Tecnologías y estándares



https://en.wikipedia.org/wiki/Semantic_Web_Stack

Web semántica y Linked Open Data

Enriquecimiento

3 owl:sameAs
skos:exactMatch
wdt:P2799

Ontologías y vocabularios

2 Vocabularios promovidos por instituciones como IFLA, W3C, DARIAH

Tecnologías y estándares

1 SPARQL, OWL, RDF, URIs

https://5stardata.info/es/

Web semántica y Linked Open Data

Cuando trabajamos con RDF, la información se almacena en forma de tripletas, en forma de Sujeto-Predicado-Objeto.

es_autor_de

Sujeto **Predicado** **Objeto**

Web semántica y Linked Open Data

Se usan **URIs para identificar a los recursos y a las propiedades**.

https://data.cervantesvirtual.com/person/40

SEGVNDA PARTE
DEL INGENIOSO
CAVALLERO
Don Quixote de la Mancha.
POR MIGVEL DE SERVANTES
Saauedra, autor de la primera Parte.
Año 1617.
EN BARCELONA, En casa de SEBASTIAN MATEVAT.
A costa de Iuan Simon mercader de Libros.

https://data.cervantesvirtual.com/manifestation/238924

Web semántica y Linked Open Data

Enriquecimiento

3

owl:sameAs
skos:exactMatch
wdt:P2799

Ontologías y vocabularios

2

Schema.org

RDA Registry

Modelo de Referencia Bibliotecaria de la IFLA

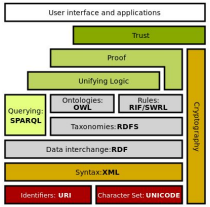
Modelo Conceptual para la Información Bibliográfica

TaDiRAH

Tecnologías y estándares

1

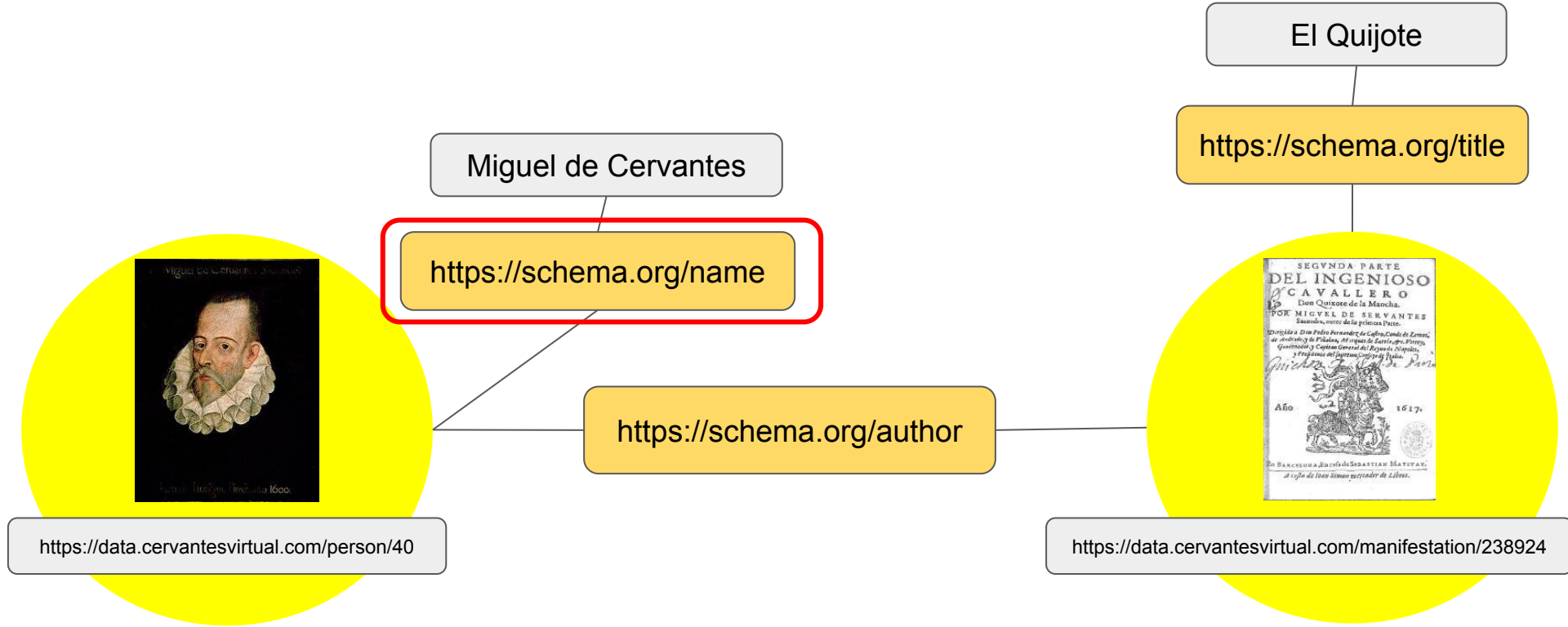
SPARQL, OWL, RDF, URIs



https://en.wikipedia.org/wiki/Semantic_Web_Stack

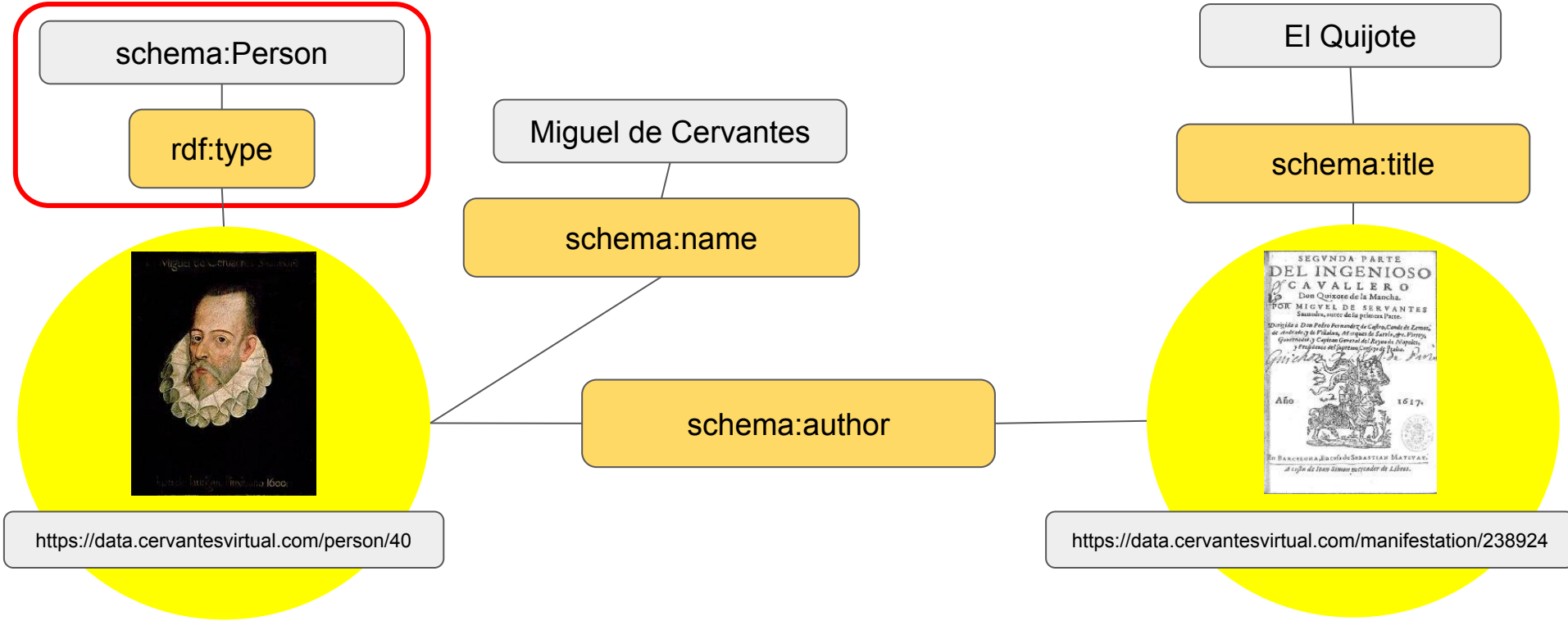
Web semántica y Linked Open Data

Podemos describir **los recursos** usando ontologías como Schema.org



Web semántica y Linked Open Data

Podemos especificar **el tipo de los recursos** y usar **espacios de nombres** (p.ej. schema en lugar de `https://schema.org/`)



Web semántica y Linked Open Data

Enriquecimiento

owl:sameAs
skos:exactMatch
wdt:P2799

3

WIKIDATA VIAF Geonames

Ontologías y vocabularios

2

Schema.org

RDA Registry

Modelo de Referencia Bibliotecaria de la IFLA

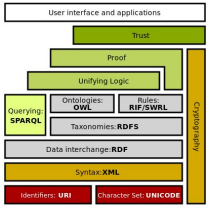
Modelo Conceptual para la Información Bibliográfica

TaDiRAH

Tecnologías y estándares

1

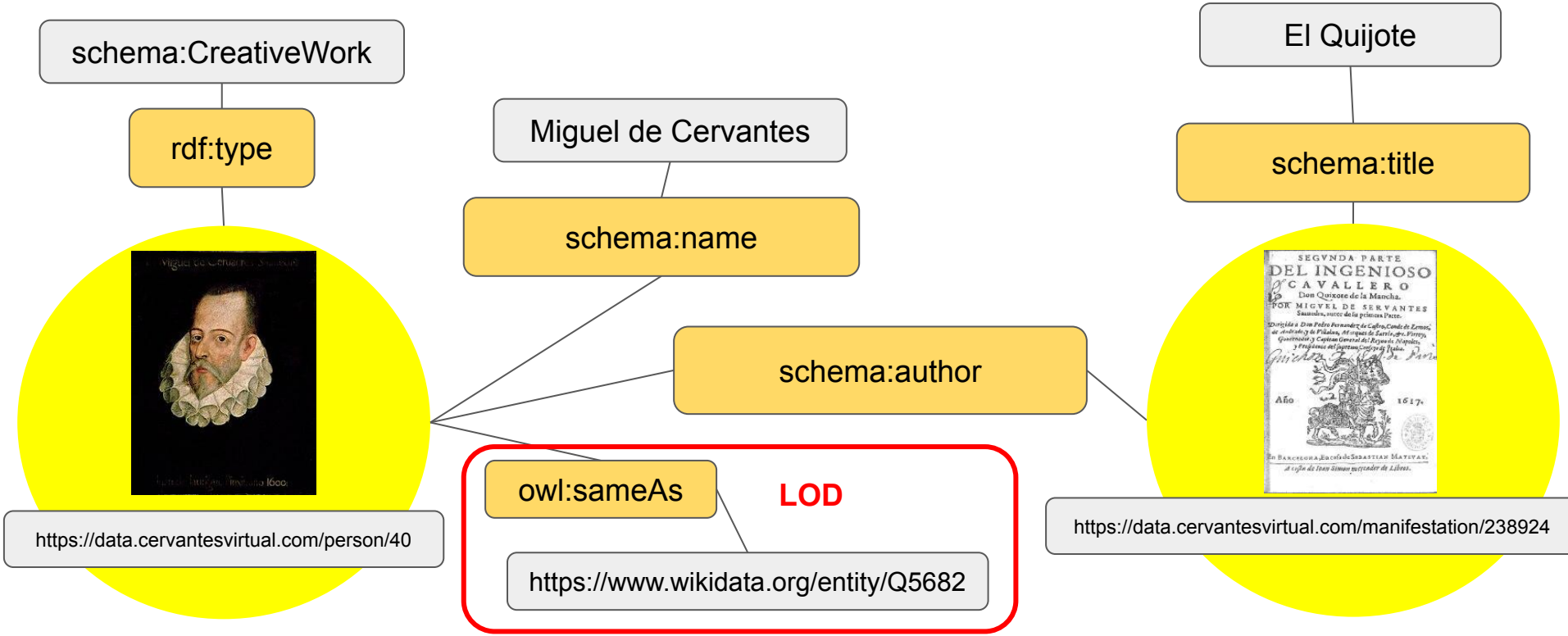
SPARQL, OWL, RDF, URIs



https://en.wikipedia.org/wiki/Semantic_Web_Stack

Web semántica y Linked Open Data

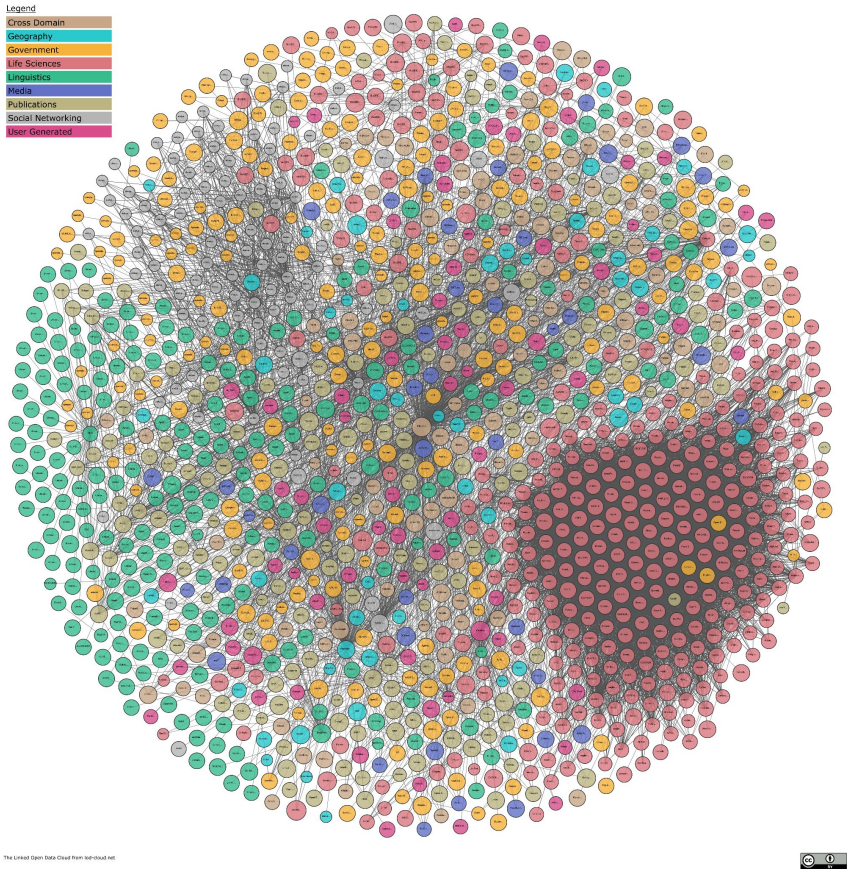
Podemos **enriquecer** los recursos con la propiedad **owl:sameAs**



Web semántica y Linked Open Data

https://lod-cloud.net/

Nov, 2025
1678 repositorios



Web semántica y Linked Open Data

europeana pro

https://pro.europeana.eu/page/apis

Data Foundry

Data collections from the National Library of Scotland

https://doi.org/10.5281/zenodo.8051036

LIBRARY LABS

LIBRARY OF CONGRESS

https://id.loc.gov

KUNGL. BIBLIOTEKET

National Library of Sweden

https://libris.kb.se/sparql

SAAM

https://americanart.si.edu/about/lod

DATOS·BNE·es

https://datos.bne.es/

{BnF Data

https://data.bnf.fr/

KB LAB

https://data.bibliotheken.nl

1640 THE NATIONAL LIBRARY OF FINLAND

https://data.nationallibrary.fi

BRITISH LIBRARY

https://bl.natbib-lod.org/

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

https://data.cervantesvirtual.com/

Bibliothèque nationale du Luxembourg

Open Data

https://data.bnl.lu/

DEUTSCHE NATIONAL BIBLIOTHEK

https://www.dnb.de/EN/lds

Österreichische Nationalbibliothek

https://labs.onb.ac.at/en/dataset/lod

Wikidata

Multilingüe

Datos estructurados

Dominio público

CC0 1.0 Universal (CC0 1.0)
Public Domain Dedication

WIKIDATA

RDF

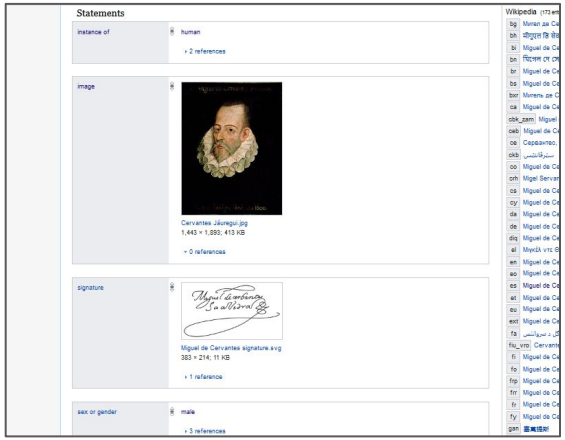
W3C SPARQL

Basado en estándares

Editado por la comunidad

Wikidata

Representación en forma de tripleta



https://www.wikidata.org/wiki/Q5682

Miguel de Cervantes nació en 1547

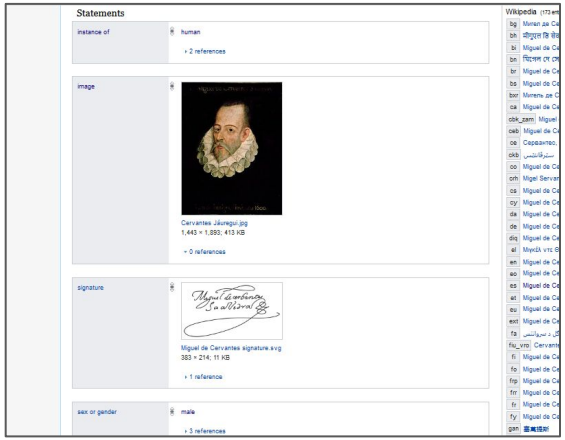
Sujeto - Predicado - Objeto

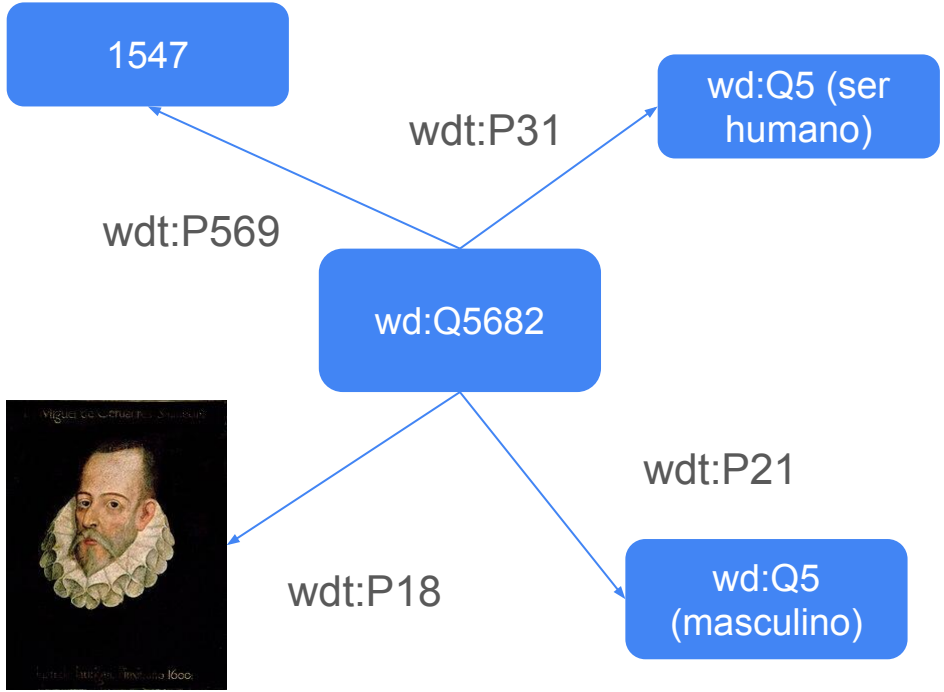
wd:Q5682 — wdt:P569 — 1547

Uso de identificadores únicos para entidades y propiedades

Wikidata

Representación en forma de grafo

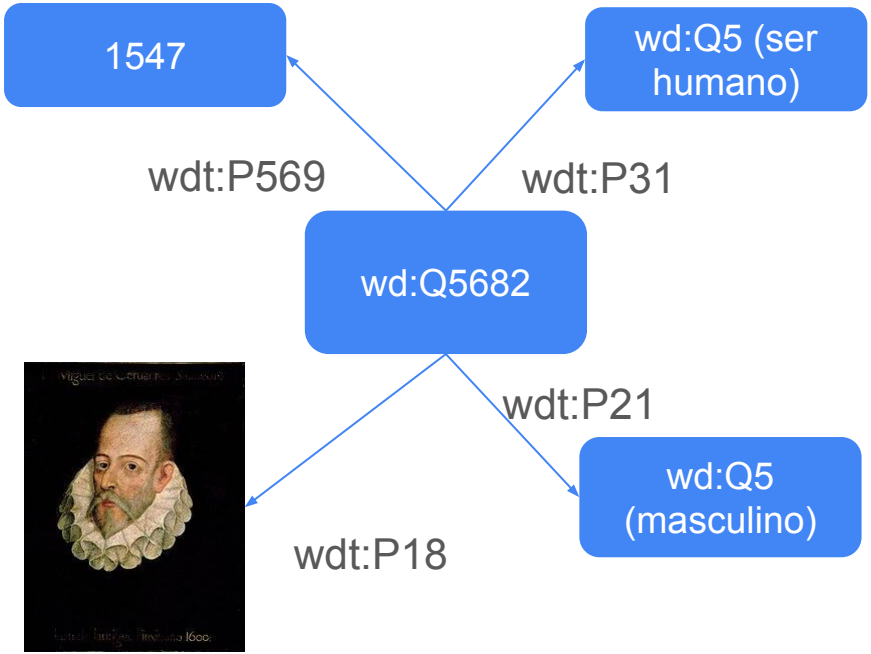




https://www.wikidata.org/wiki/Q5682

Wikidata

Espacios de nombre



wd = http://www.wikidata.org/entity/
wd:Q5682 = http://www.wikidata.org/entity/Q5682

wdt = http://www.wikidata.org/prop/direct/
wdt:P31 = http://www.wikidata.org/prop/direct/P31

Podemos usar los prefijos wd: y wdt: en lugar de la URL completa

https://www.mediawiki.org/wiki/Wikibase/Indexing/RDF_Dump_Format#Full_list_of_prefixes

Consultas con SPARQL

Introducción

Espacios de nombre a usar en la sentencia

```
PREFIX ex: <example.org>
SELECT *
WHERE {
    ?s ?p ?o
}
```

Selección de columnas. Un * se refiere a todas

Tripletas para restricciones. Las que empiezan con “?” las interpreta como variables

Consultas con SPARQL

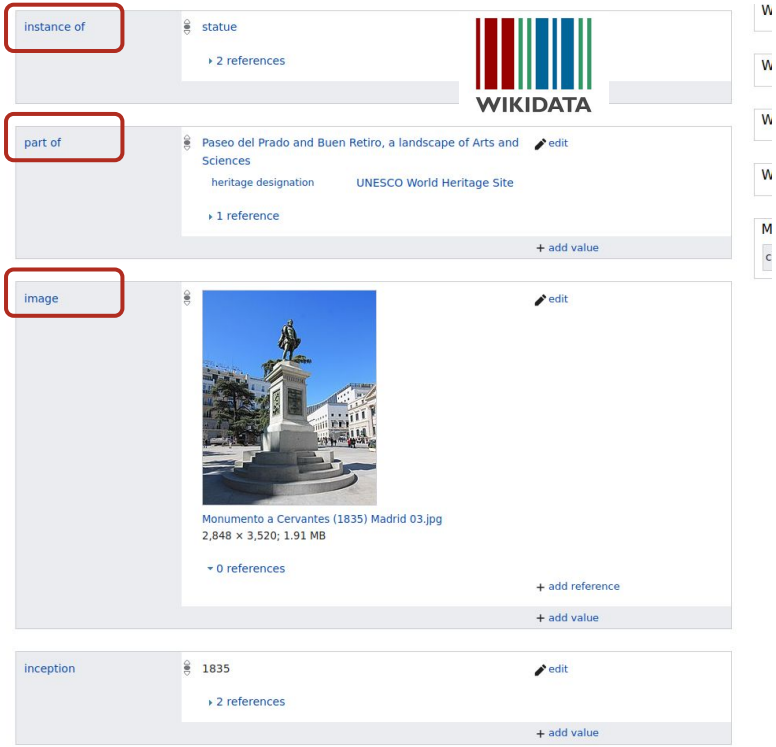
¿Qué estructura tienen los datos?

- Revisar las propiedades utilizadas para describir los metadatos
- Obtener las clases y propiedades mediante sentencias SPARQL

```
SELECT DISTINCT ?class
WHERE {
  ?subject a ?class .

  FILTER( REGEX (?class, "rdaregistry"))
}
LIMIT 100
```

https://data.cervantesvirtual.com/sparql

https://datos.bne.es/sparql



https://www.wikidata.org/wiki/Wikidata:List_of_properties

Consultas con SPARQL

Introducción

https://w.wiki/88yd

?autor devuelve el identificador del recurso en Wikidata

```
SELECT ?autor ?idbvmc
WHERE {
    ?autor wdt:P2799 ?idbvmc
}
LIMIT 10
```

Propiedad que enlaza autores de la BVMC en Wikidata

Consultas con SPARQL

Introducción

Si añadimos **Label** al nombre de la variable automáticamente nos recupera la etiqueta del recurso. ?autor devuelve el identificador del recurso

https://w.wiki/85wT

```
SELECT ?autor ?autorLabel ?idbvmc
WHERE {
    ?autor wdt:P2799 ?idbvmc .

    SERVICE wikibase:label {
        bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es".}
}
LIMIT 10
```

Añadimos esta instrucción para poder tener acceso a las etiquetas de forma más sencilla. AUTO_LANGUAGE se refiere al idioma del navegador

Consultas con SPARQL

Introducción

https://w.wiki/89LA

```
SELECT ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento

  FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
    && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime).

  SERVICE wikibase:label {
     bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es".}
}
LIMIT 10
```

Instrucción FILTER permite aplicar filtros

Consultas con SPARQL

Uso avanzado

[Más ejemplos de SPARQL]

- Operadores Values y Bind
  - asignar uno o más valores específicos a una variable
  - asignar el resultado de una expresión a una variable
- Service
  - consultas federadas entre repositorios de datos SPARQL
- Visualizaciones en Wikidata
  - #defaultView:Map, #defaultView:Graph
- Expresiones regulares
  - FILTER regex(?x, "pattern" [, "flags"])
  - FILTER regex(?nombre, "g", "i"): la variable ?nombre contiene el texto “g” o “G”
- Limit y offset: paginación resultados

Ontologías y vocabularios

¿Qué es una ontología?

ontología

Artículo | Sinónimos o afines

Del lat. mod. *ontologia*, de *onto-* 'onto-' y *-logia* '-logía'.

Diccionario de la lengua española

1. f. *Fil.* Parte de la metafísica que trata del ser en general y de sus propiedades trascendentales.

   SIN.: **metafísica**, **filosofía**.

2. f. En ciencias de la comunicación y en inteligencia artificial, red o sistema de datos que define las relaciones existentes entre los conceptos de un dominio o área del conocimiento.

Clases y propiedades para describir los recursos

Ontologías y vocabularios

Ontologías para modelar y describir la información

Schema.org

https://schema.org/

bf: BIBFRAME

https://www.loc.gov/bibframe/

europeana

EUROPEANA DATA MODEL

https://pro.europeana.eu/page/edm-documentation

RDA Registry

http://www.rdaregistry.info/

IFLA International Federation of Library Associations and Institutions

IFLA Library Reference Model

https://repository.ifla.org/handle/123456789/40

Data Catalog Vocabulary (DCAT)

https://www.w3.org/TR/vocab-dcat-3/

CIDOC CRM CONCEPTUAL REFERENCE MODEL

https://www.cidoc-crm.org/

Ontologías y vocabularios

¿Qué es una vocabulario o taxonomía?

taxonomía

Artículo | Sinónimos o afines

Del gr. τάξις *táxis* 'ordenación' y *-nomía*.

1. f. Ciencia que trata de los principios, métodos y fines de la clasificación. Se aplica en particular, dentro jerarquizada y sistemática, con sus nombres, de los grupos de animales y de vegetales.

2. f. **clasificación** (|| acción de clasificar).

   SIN.: **clasificación**, **categorización**, **sistemática**.

Diccionario de la lengua española

Términos para describir y clasificar los recursos

Ontologías y vocabularios

TADIRAH

THE TAXONOMY OF DIGITAL RESEARCH ACTIVITIES IN THE HUMANITIES

skos.um.es/unesco6/00/html

SKOS

Nomenclatura de Ciencia y Tecnología de la UNESCO

- Introducción
- Consultar nomenclatura
- Presentación alfabética
- Presentación jerárquica
- Punto de acceso SPARQL
- Descargas
- Estadísticas
- Créditos y aviso legal

Español | English | Français | Texto a buscar

RDF/XML | N-Triples | N3/Turtle | JSON | JSON-LD

Proyecto de nomenclatura internacional normalizada relativa a la ciencia y la tecnología

Proposed international standard nomenclature for fields of science and technology **(en)**
Projet de nomenclature internationale type des domaines de la science et de la technologie **(fr)**

Conceptos cabecera (skos:topConcept)

- 11 Lógica
- 12 Matemáticas
- 21 Astronomía y Astrofísica
- 22 Física
- 23 Química
- 24 Ciencias de la Vida
- 25 Ciencias de la Tierra y del Espacio
- 31 Ciencias Agrarias
- 32 Ciencias Médicas
- 33 Ciencias Tecnológicas
- 51 Antropología
- 52 Demografía
- 53 Ciencias Económicas
- 54 Geografía
- 55 Historia
- 56 Ciencias Jurídicas y Derecho
- 57 Lingüística
- 58 Pedagogía
- 59 Ciencia Política
- 61 Psicología
- 62 Ciencias de las Artes y las Letras
- 63 Sociología
- 71 Ética
- 72 Filosofía

The LIBRARY of CONGRESS | ASK A LIBRARIAN | DIGITAL COLLECTIONS | LIBRARY CATALOGS | Search | GO

The Library of Congress > Cataloging, Acquisitions > Library of Congress Subject Headings PDF Files

CATALOGING AND ACQUISITIONS

- Cataloging and Acquisitions Home
- About the Organization
- Contact
- FAQs
- News
- Acquisitions
- Resources for Cataloging
- Catalogs, Authority Records
- Classification and Shelflisting
- Cooperative Cataloging Programs
- Descriptive Cataloging
- Products for Purchase
- Publications, Reports
- Subject Headings & Genre/Form Terms

Library of Congress Subject Headings PDF Files

About LCSH - Introduction to LCSH - Individual Files - Free-floating Subdivisions - Children's Subject Headings - Genre/Form Terms for Library and Archival Materials - Demographic Group Terms - Medium of Performance Thesaurus for Music -

About the Library of Congress Subject Headings PDF Files

This page provides print-ready PDF files for the 43rd Edition of the Library of Congress Subject Headings (LCSH). Data for the 43rd edition was selected in April 2021. For users desiring enhanced functionality, LCSH will continue as part of the web-based subscription product, *Classification Web*.

Earlier editions are available here but should not be used for cataloging.

Back to Top

Introduction to Library of Congress Subject Headings

Introduction to Library of Congress Subject Headings *(PDF, 11 p., 337 KB)*

Back to Top

Individual PDF Files: By Beginning Letter

Numerals (12 p., 143 KB)	A (563 p., 3.14 MB)	B (490 p., 2.74 MB)
C (842 p., 4.65 MB)	D (328 p., 1.86 MB)	E (323 p., 1.84 MB)
F (322 p., 1.84 MB)	G (323 p., 1.83 MB)	H (334 p., 1.89 MB)

European Commission | EN | Search

Live, work, travel in the EU

Employment, Social Affairs and Inclusion

Home | Policies and activities | News | Events | Publications | Contact | #EUSocialForum

Employment, Social Affairs and Inclusion > ... > Skills for jobs > European Skills/Competences, Qualifications and Occupations (ESCO)

European Skills/Competences, Qualifications and Occupations (ESCO)

ESCO (European Skills, Competences, Qualifications and Occupations) is the European multilingual classification of skills, competences, qualifications and occupations.

ESCO works like a dictionary, describing, identifying and classifying professional occupations and skills relevant for the EU labour market and education and training area and systematically showing the relations between those occupations and skills. It is available in an online portal where its dataset of occupations and skills can be consulted and downloaded free of charge.

Its **common reference terminology** helps make the European labour market more effective and integrated, and allows the worlds of work and education/training to communicate more effectively with each other.

What is ESCO?

Schema.org

Schema.org Docs Schemas Validate About

Schema.org is a collaborative, community activity with a mission to create, maintain, and promote schemas for structured data on the Internet, on web pages, in email messages, and beyond.

Schema.org vocabulary can be used with many different encodings, including RDFa, Microdata and JSON-LD. These vocabularies cover entities, relationships between entities and actions, and can easily be extended through a well-documented extension model. As of 2024, over 45 million web domains markup their web pages with over 450 billion Schema.org objects. Many applications from Google, Microsoft, Pinterest, Yandex and others already use these vocabularies to power rich, extensible experiences.

Founded by Google, Microsoft, Yahoo and Yandex, Schema.org vocabularies are developed by an open community process, using the public-schemaorg@w3.org mailing list and through GitHub.

Place

A Schema.org Type

Thing › Place

[more...]

Entities that have a somewhat fixed, physical extension.

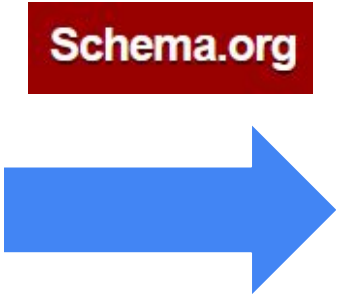
Property	Expected Type	Description
Properties from Place		
additionalProperty	PropertyValue	A property-value pair representing an additional characteristic of the entity, e.g. a product feature or another characteristic for which there is no matching property in schema.org. Note: Publishers should be aware that applications designed to use specific schema.org properties (e.g. https://schema.org/width, https://schema.org/color, https://schema.org/gtin13, ...) will typically expect such data to be provided using those properties, rather than using the generic property/value mechanism.
address	PostalAddress or Text	Physical address of the item.
aggregateRating	AggregateRating	The overall rating, based on a collection of reviews or ratings, of the item.
amenityFeature	LocationFeatureSpecification	An amenity feature (e.g. a characteristic or service) of the Accommodation. This generic property does not make a statement about whether the feature is included in an offer for the main accommodation or available at extra costs.
branchCode	Text	A short textual code (also called "store code") that uniquely identifies a place of business. The code is typically assigned by the parentOrganization and used in structured URLs. For example, in the URL http://www.starbucks.co.uk/store-locator/etc/detail/3047 the code "3047" is a branchCode for a particular branch.
containedInPlace	Place	The basic containment relation between a place and one that contains it. Supersedes containedIn. Inverse property: containsPlace
containsPlace	Place	The basic containment relation between a place and another that it contains. Inverse property: containedInPlace
event	Event	Upcoming or past event associated with this place, organization, or action. Supersedes events.

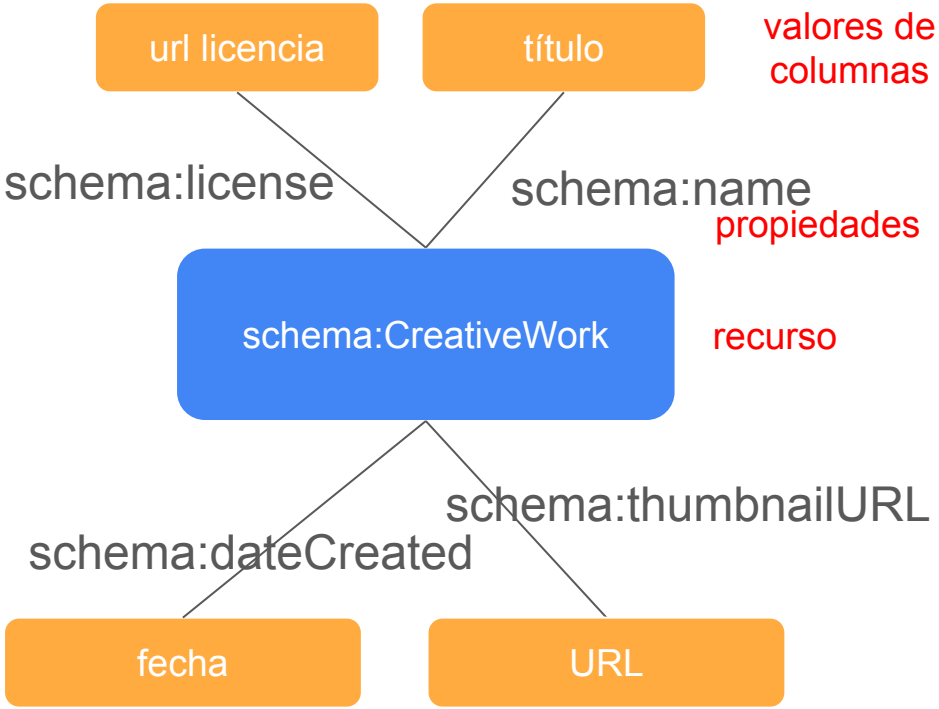
https://schema.org/Place

Ontologías

CSV metadatos:

- título
- licencia
- imagen
- fecha





Ontologías

The KnowWhereGraph ontology

https://knowwheregraph.org/

At the project's core is the KnowWhere Graph, a geo-knowledge graph that is based on existing standards like RDF, OWL and GeoSPARQL, incorporates custom ontologies, and uses a hierarchical grid for spatial representations. The integrated KWG schema provides a holistic view of the graph modeling. Its current size exceeds 12 billion information triples, and the covered data support pilot scenarios in disaster relief, agricultural land use and food-related supply chains. These data include observations of natural hazards (e.g., hurricanes, wildfires, smoke plumes) and spatial characteristics related to climate (e.g., temperature, precipitation, air quality), soil properties, crop and land-cover types, demographics, and human health, among others. The table below shows a summary of the raw datasets, their providers, their spatiotemporal scopes etc.

Thematic Datasets					Place-Centric Datasets		
Dataset Name/ Theme	Source Agency	Key Attributes	Spatial Coverage	Temporal Coverage	Place-Centric Dataset	Defining Authority	Spatial Coverage
Soil Properties	USDA	soil type, farmland class	Targeted regions in US	Current	S2 Cells	Google	Lvl 9 (Global), Lvl 13 (US),
Wildfires	USGS, USDA, USFS, NIFC	wildfire type, burn severity, num. acres burned, contained date	US	1984–current	Global Administrative Regions	University of Berkeley, Museum of Vertebrate Zoology and the International	Global
Earthquakes	USGS	magnitude, length, width, geometry	Global (mag. over 4.5)	2011-01-01 to 2022-01-18			

Ontologías

GeoNames Home | Postal Codes | Download / Webservice | About

Alicante | all countries

search [advanced search]

2802 records found for "Alicante"

	Name	Country	Feature class	Latitude	Longitude
1	Alicante A-li-kham-thit,ALC,Akra Leuke,Alacant,Alacante,Alacanti,Alakanto,Alicant,Alicante,Alikante,Alikanteh...	Spain, Valencia Alicante > Alicante	seat of a second-order administrative division population 348,901	N 38° 20′ 42″	W 0° 28′ 53″
2	Elche Ehl'ch,Ehl'che,El'che,Elce,Elch,Elche,Elig,Elx,Elče,Ilici,ai er qie,alchh,alsh,elche,eruche,xelche,...	Spain, Valencia Alicante > Elche	seat of a third-order administrative division population 234,765, elevation 96m	N 38° 15′ 43″	W 0° 42′ 3″
3	Alicante-Elche Airport ALC,Aeroport d'Alicante-Elche,Aeroporto di Alicante-Elche,Aeropuerto Internacional de Alicante - El ...	Spain, Valencia Alicante > Elche	airport elevation 43m	N 38° 16′ 55″	W 0° 33′ 29″
4	Alicante Canyon		canyon	N 37° 59′ 0″	E 0° 6′ 0″
5	Orihuela Auraiola,Aurariola,Orihuela,Oriola,Oriouela,Oriuehla,Oriuela,Orivela,ao li wei la,aryhwyla,awrywyla,...	Spain, Valencia Alicante > Orihuela	seat of a third-order administrative division population 101,321	N 38° 5′ 5″	W 0° 56′ 38″
6	Torrevieja Malnovturo,Toreviekha,Toreviexa,Torevijekha,Torrev'ekha,Torrevekha,Torrevella,Torrevella de la Mata,...	Spain, Valencia Alicante > Torrevieja	seat of a third-order administrative division population 82,599	N 37° 58′ 43″	W 0° 40′ 56″
7	Benidorm Benidorm,Benidormo,bei ni duo er mu,benidoleum,benidorumu,bnydwrm,Бенидорм,Бенідорм,Բենիդորմ,بنيدورم...	Spain, Valencia Alicante > Benidorm	seat of a third-order administrative division population 70,450	N 38° 32′ 17″	W 0° 7′ 51″
8	Alcoy Al'ka,Al'koj,Alco,Alcodium,Alcoi,Alcoy,Alkoj,Alkojus,a er ke yi,alkwy,arukoi,Алкоj,Алька,Алькой,آلكوي...	Spain, Valencia Alicante > Alcoy	seat of a third-order administrative division population 61,552	N 38° 42′ 19″	W 0° 28′ 27″

GeoNames

https://www.geonames.org/

Ontologías

Creando nuestra propia ontología….

https://protege.stanford.edu/

Transformación

Tecnologías

- Apache Jena (RDF Java framework)
- Virtuoso open-source edition
- Spring Boot application
- Apache Maven
- RDFLib
- Jupyter
- Neo4J

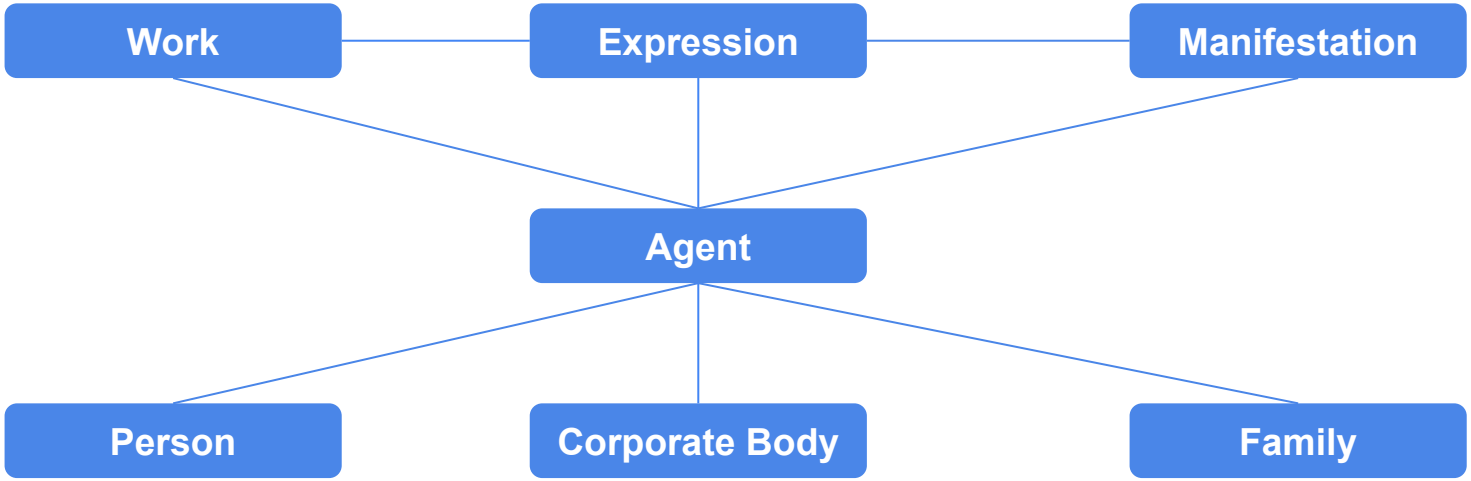
Transformación

En 2015 la BVMC publicó su catálogo como datos abiertos y enlazados

- Modelado de datos basado en RDF y **RDA registry**
- Enriquecimiento con repositorios externos como **Wikidata y GeoNames**
- Disponible a través de un punto de acceso SPARQL
- Candela, G., Escobar, P., Carrasco, R. and Marco-Such, M. Migration of a library catalogue into RDA linked open data. Semantic Web 9(4): 481-491 (2018). http://hdl.handle.net/10045/65427

Transformación

Modelado de datos



RDA Registry

http://www.rdaregistry.info/Elements/c/

Transformación

Varios puntos de acceso (URIs) usados como **identificadores únicos**:

- https://data.cervantesvirtual.com/**date**/1616
- https://data.cervantesvirtual.com/**language**/es
- https://data.cervantesvirtual.com/**manifestation**/722875
- https://data.cervantesvirtual.com/**person**/40
- https://data.cervantesvirtual.com/**corporatebody**/7802
- https://data.cervantesvirtual.com/**work**/2904
- https://data.cervantesvirtual.com/**expression**/152342
- https://data.cervantesvirtual.com/**item**/134373

Nuevas formas de acceso al catálogo

Transformación

Wikidata y la Biblioteca Virtual Miguel de Cervantes

Propiedad	Descripción	Elementos
P2799	BVMC person ID	15054
P3976	BVMC work ID	764
P10834	BVMC organization ID	240

https://www.wikidata.org/wiki/Wikidata:Property_proposal

Transformación

Wikidata y la Biblioteca Virtual Miguel de Cervantes

Vega, Lope de, 1562-1635

Wikimedia Commons

sacerdote, poeta y dramaturgo español del Siglo de Oro

Información extraída de Wikipedia / (CC BY-SA 3.0)

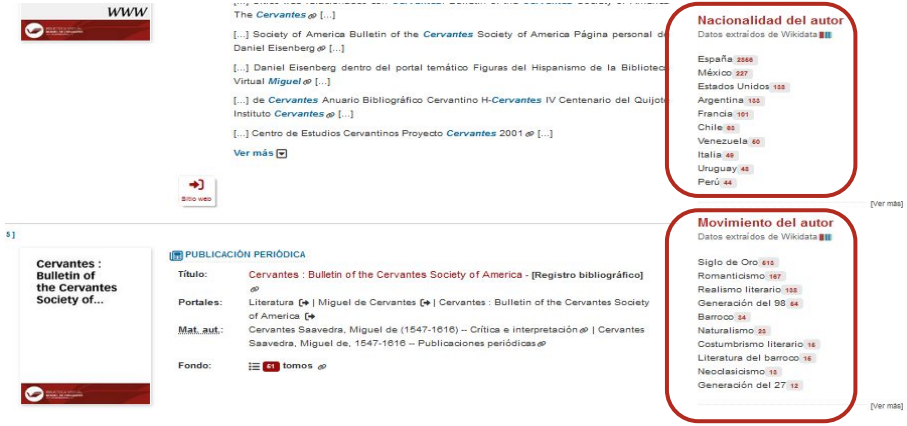
Identificador: 72
Nombre: Vega, Lope de
Fecha de nacimiento: 1562
Fecha de fallecimiento: 1635
VIAF: https://viaf.org/viaf/89773778
Variantes de nombre:
- Burguillos, Tomé de
- Carrera, Luis de la
- Flórez, Antonio
- Lope de Vega Carpio, Félix
- Padecopio, Gabriel
- Puente, Juan de la
- Un Ingenio de esta Corte
- Vega Carpio, Lope de
- Vega y Carpio, Lope de
- Lope de Vega

Lugar de nacimiento: Madrid
Lugar de fallecimiento: Madrid
Nacionalidad: España
Nombre de pila: Lope
Género: masculino

EXPORTAR:

RDF JSON

https://data.cervantesvirtual.com/person/72



https://www.cervantesvirtual.com/buscador/?q=miguel+de+cervantes

Transformación - Java y Jena

Proyecto biblioteca de Spring Boot en version 3.4.2

Este proyecto se ha creado para la asignatura Tecnologías de Internet Orientadas al Navegador del Máster Universitario en Desarrollo de Aplicaciones y Servicios Web de la Universidad de Alicante.

Para poder editar el código es posible installar un IDE como Eclipse o Idea.

Configuración

En el fichero `src/main/resources/application.properties` se puede configurar el acceso a la base de datos mysql para el almacenamiento de los libros del catálogo. También se puede configurar el puerto para ejecutar la aplicación.

El proyecto incluye la librería Jena como dependencia en el fichero `pom.xml` para el uso de SPARQL.

El fichero `pom.xml` incluye las versiones de las librerías utilizadas en este proyecto.

API REST

El proyecto proporciona los siguientes patrones de URL para dar acceso al API REST:

- Crear libro (POST): localhost:8081/api/book
- Modificar libro (UPDATE): localhost:8081/api/book/{ID}
- Eliminar libro (DELETE): localhost:8081/api/book/{ID}
- Recuperar libro (GET): localhost:8081/api/book/{ID}

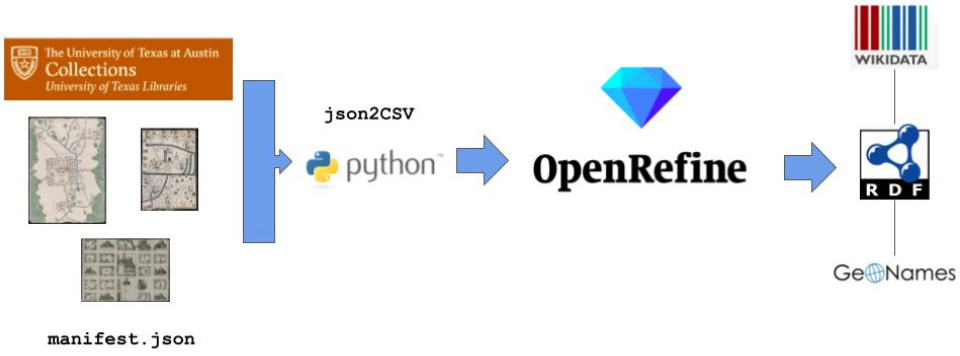
Para poder realizar pruebas se puede utilizar la herramienta Postman

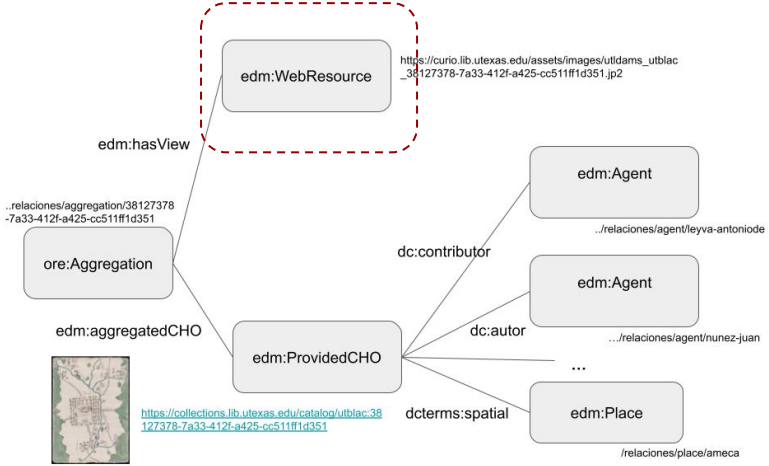
```xml
<dependencies>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-data-jpa</artifactId>
    </dependency>
    <dependency>
        <groupId>org.springframework.boot</groupId>
        <artifactId>spring-boot-starter-web</artifactId>
    </dependency>
    <dependency>
      <groupId>org.springframework.boot</groupId>
      <artifactId>spring-boot-starter-thymeleaf</artifactId>
    </dependency>

    <!-- https://mvnrepository.com/artifact/com.itextpdf/itextpdf -->
    <dependency>
        <groupId>com.itextpdf</groupId>
        <artifactId>itextpdf</artifactId>
        <version>5.5.13.4</version>
    </dependency>

    <!-- jeva version 4.1.0 Java 11 -->
    <dependency>
        <groupId>org.apache.jena</groupId>
        <artifactId>apache-jena-libs</artifactId>
        <version>5.3.0</version>
        <type>pom</type>
    </dependency>
```

https://github.com/hibernator11/biblioteca-springboot

Transformación

Unlocking the Colonial Archive



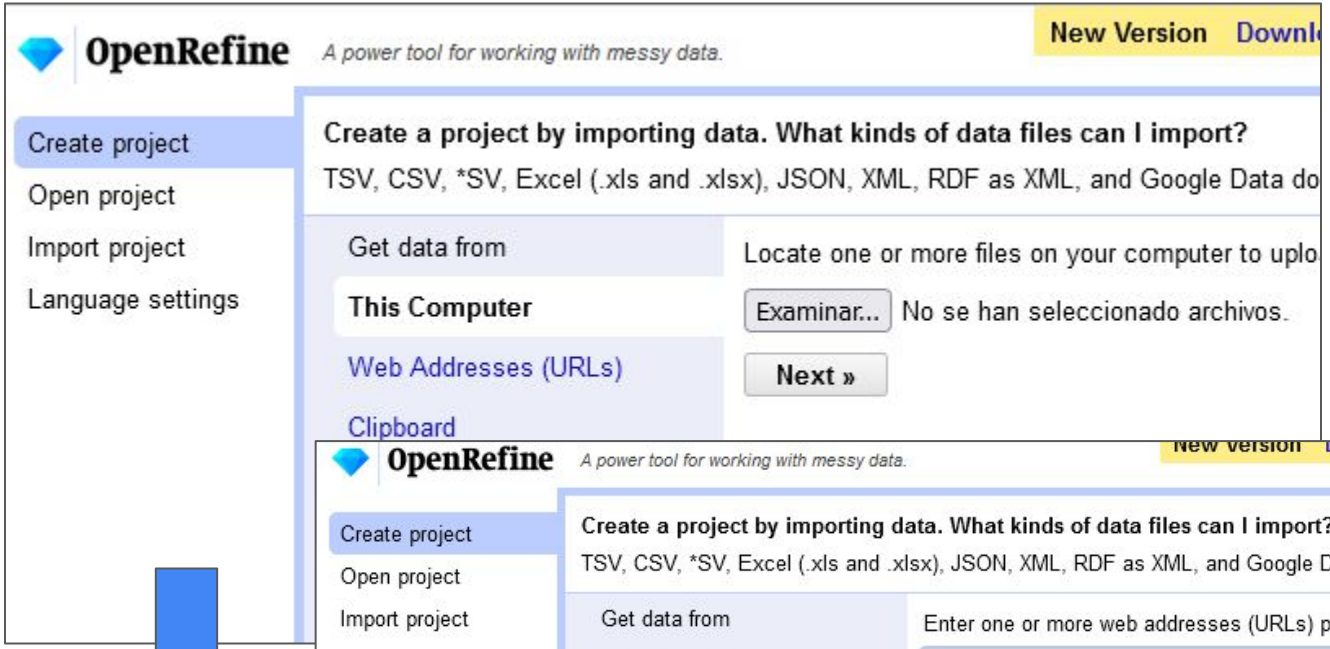


Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2023. An ontological approach for unlocking the Colonial Archive. J. Comput. Cult. Herit. Just Accepted (April 2023). https://doi.org/10.1145/3594727

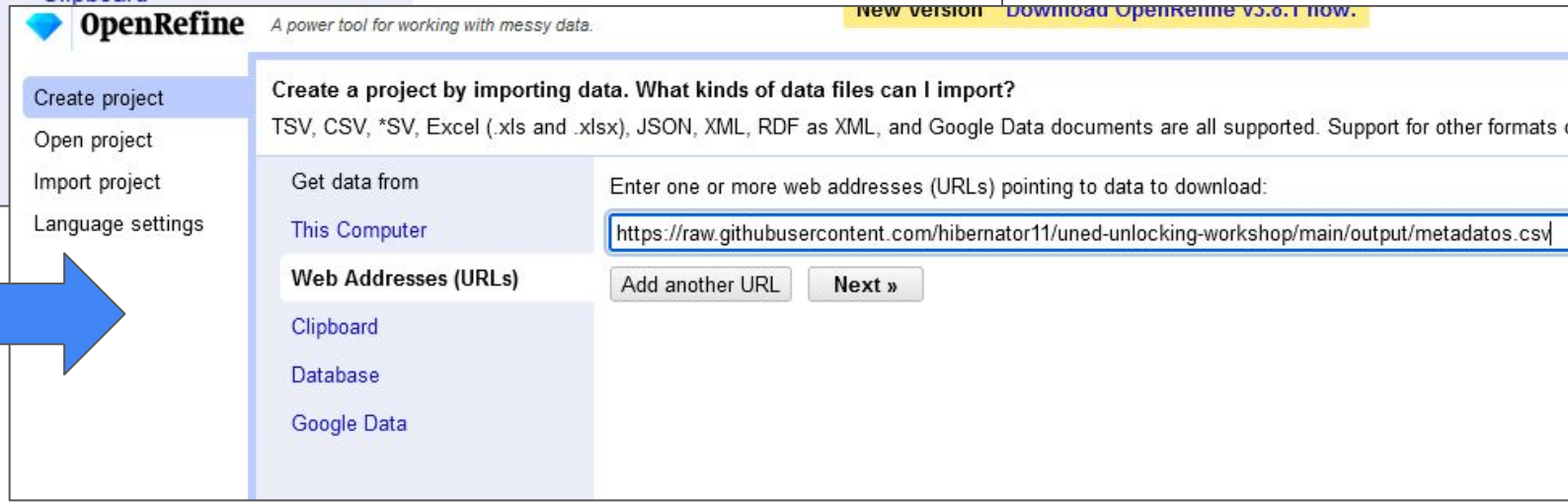
https://github.com/hibernator11/UCA-relacionesgeograficas

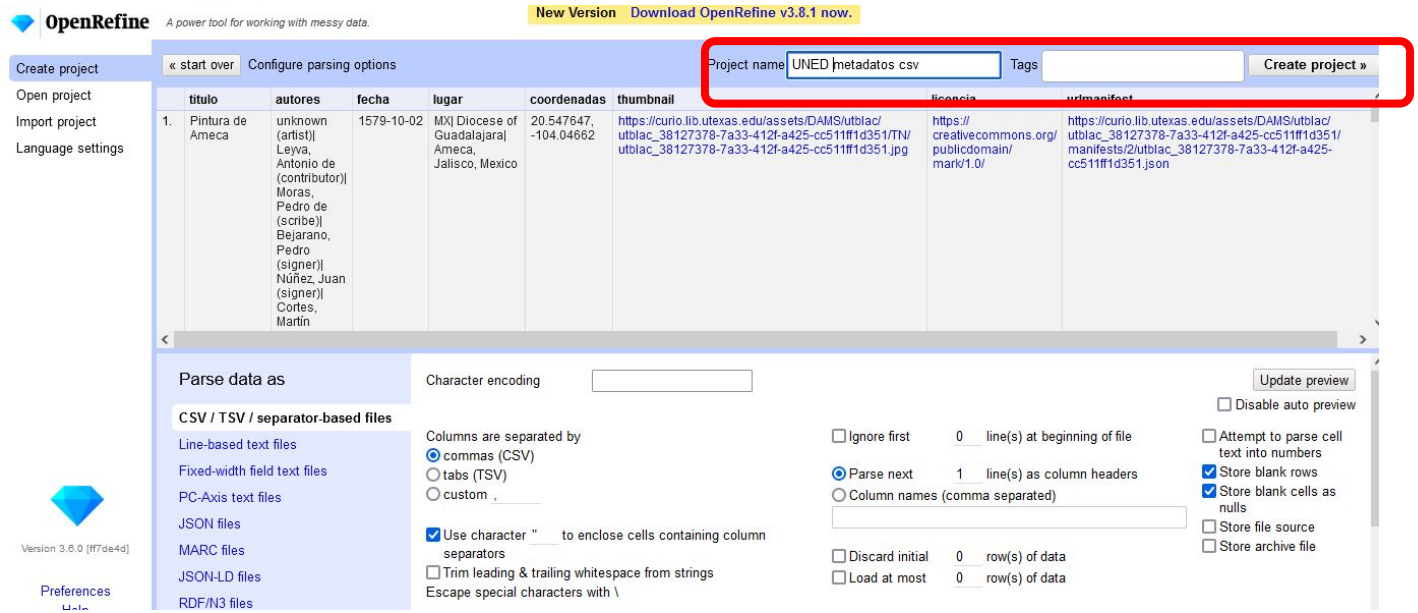
Transformación

https://github.com/hibernator11/uned-unlocking-workshop



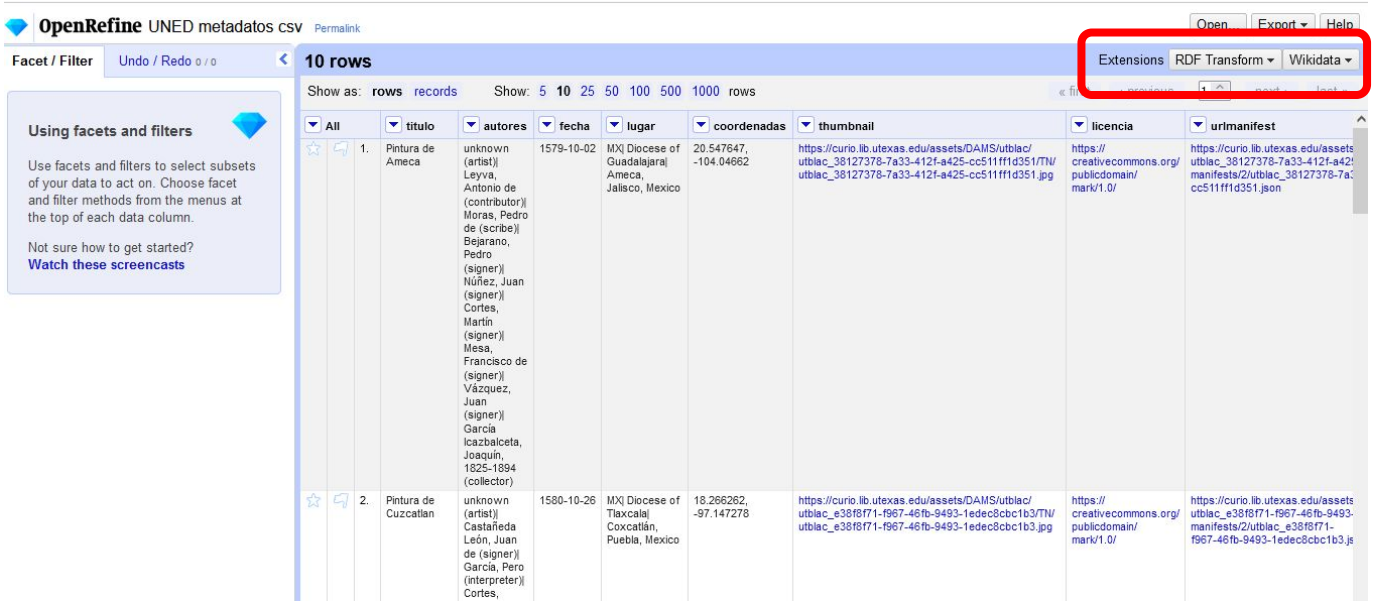
Abrimos el fichero CSV



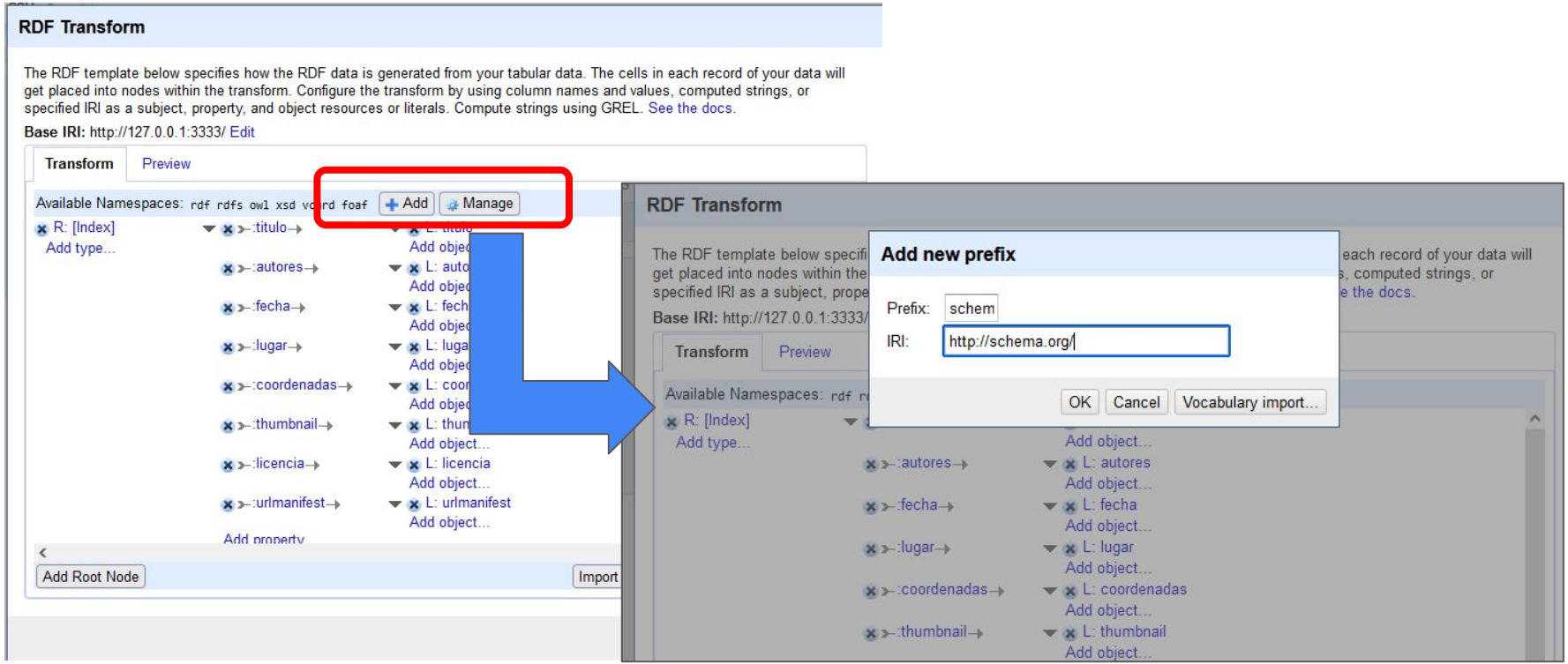


Le damos un nombre al proyecto

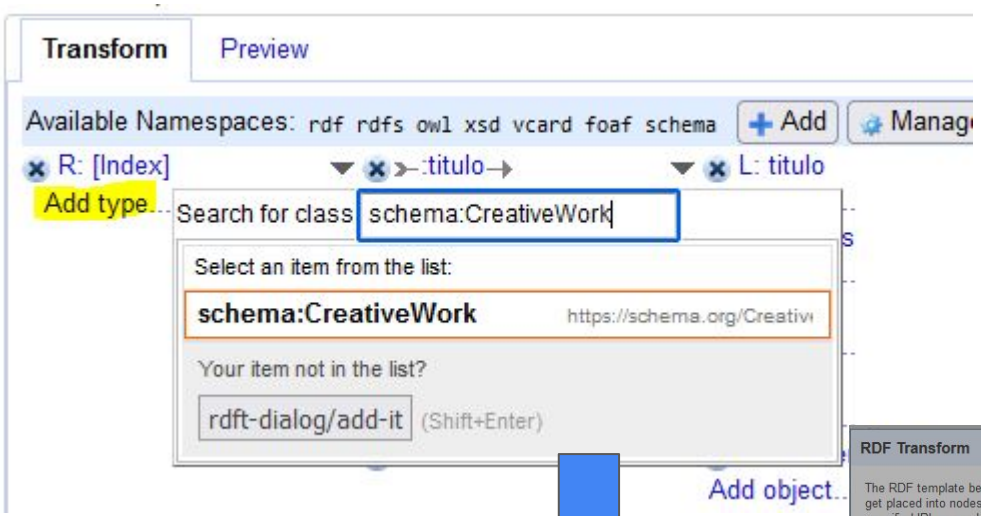
Le damos un nombre y creamos el proyecto desde el menú de la parte superior



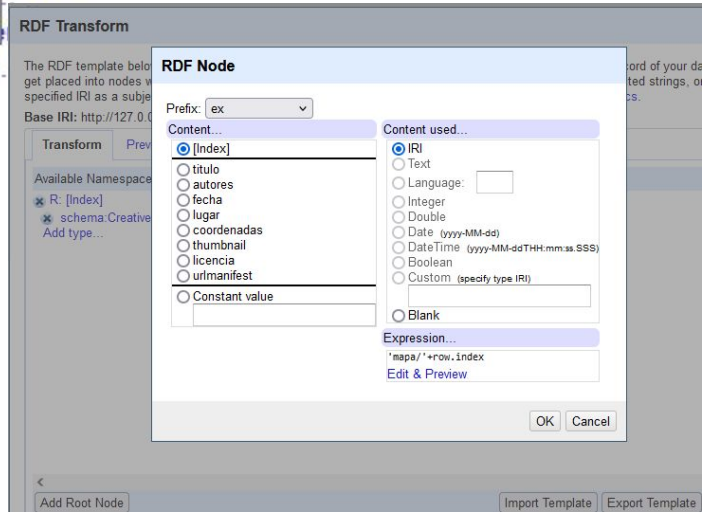
Vista de creación del proyecto, desde el menú superior podemos configurar la extensión para generar el formato RDF

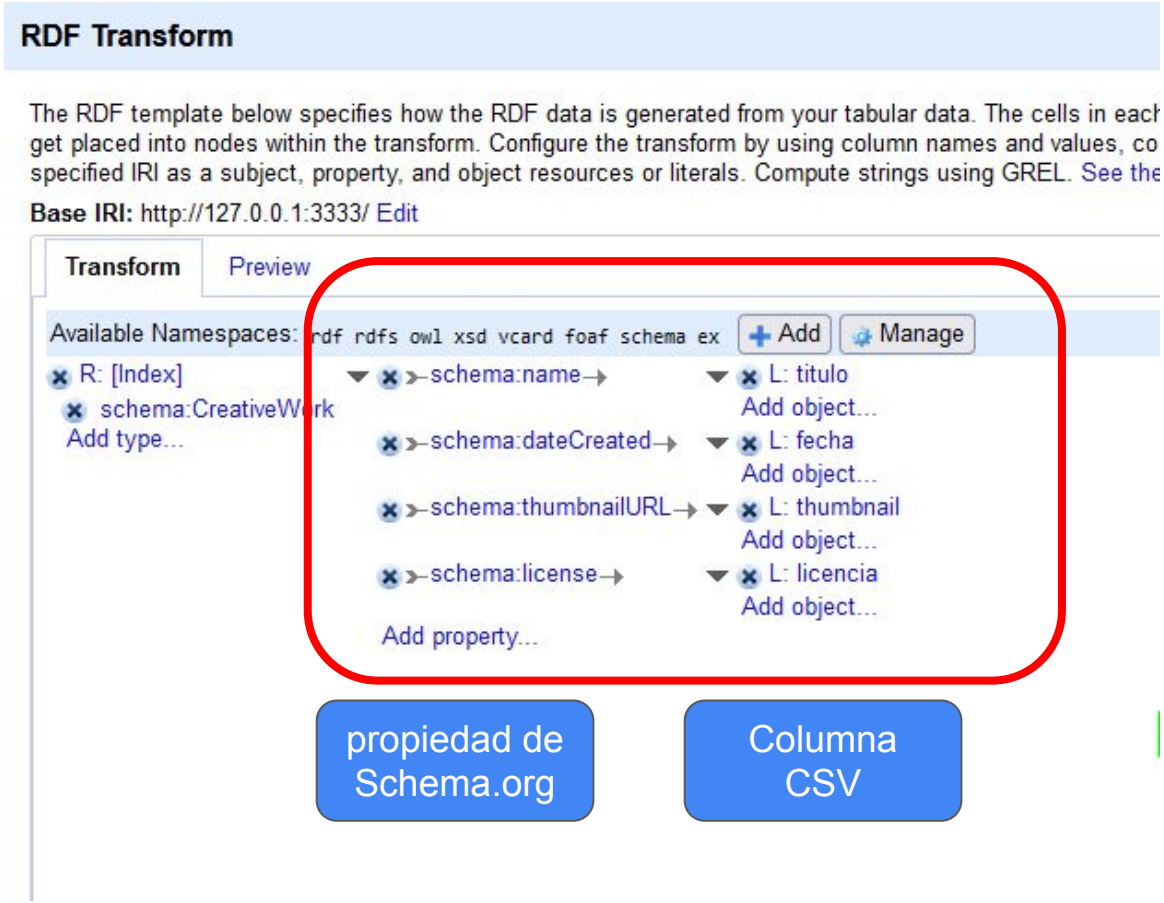


Añadimos el espacio de nombres Schema.org y https://example/ para crear las URIs (http://example/mapa/{xyz})



Añadimos la clase y definimos la URL





Añadimos las propiedades

RDF Transform

The RDF template below specifies how the RDF data is generated from your tabular data. The cells in each record of your data w
get placed into nodes within the transform. Configure the transform by using column names and values, computed strings, or
specified IRI as a subject, property, and object resources or literals. Compute strings using GREL. See the docs.

Base IRI: http://127.0.0.1:3333/ Edit

Transform **Preview**

Sample Data Preview: 20 rows. Shown below: Turtle (Stream) Turtle (Pretty)

```
@prefix schema: <https://schema.org/> .
@prefix vcard:  <http://www.w3.org/2006/vcard/ns#> .
@prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .

<http://example/mapa/3>
        rdf:type              schema:CreativeWork ;
        schema:dateCreated    "1579-11-22" ;
        schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
        schema:name           "Pintura de Zapotitlan" ;
        schema:thumbnailURL   "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_32713c87-b23c-474c-
acb9-0757bc1d2dd3/TN/utblac_32713c87-b23c-474c-acb9-0757bc1d2dd3.jpg" .

<http://example/mapa/8>
        rdf:type              schema:CreativeWork ;
        schema:dateCreated    "1580-04-29" ;
        schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
        schema:name           "Pintura de Coatzacualco" ;
        schema:thumbnailURL   "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_9a78720b-d927-4635-
bdc0-9993a3024f9a/TN/utblac_9a78720b-d927-4635-bdc0-9993a3024f9a.jpg" .

<http://example/mapa/1>
        rdf:type              schema:CreativeWork ;
```

Podemos ver el resultado desde la opción Preview

Open... Export

Show: 5 10 25 50 100 500 1000 rows

autores	fecha	lugar	coordenadas	thumbnail
:nown :ist)| rva, :onio de ntributor)| ras, Pedro (scribe)| arano, lro ner)| iez, Juan ner)| tes, rtín ner)| sa, ncisco de ner)| zquez, n ner)| rcía zbalceta, quín, 5-1894 llector)	1579-10-02	MX| Diocese of Guadalajara| Ameca, Jalisco, Mexico	20.547647, -104.04662	https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_38127378-7a33-412f-a425-cc511ff1d351/TN/utblac_38127378-7a33-412f-a425-cc511ff1d351.jpg
:nown :ist)| stañeda n, Juan (signer)| rcía, Pero erpreter)| tes,	1580-10-26	MX| Diocese of Tlaxcala| Coxcatlán, Puebla, Mexico	18.266262, -97.147278	utblac_e38f8f71-f967-46fb-9493-1edec8cbc b3.jpg

- OpenRefine project archive to file
- Tab-separated value
- Comma-separated value
- HTML table
- Excel (.xls)
- Excel 2007+ (.xlsx)
- ODF spreadsheet
- Custom tabular exporter...
- SQL Exporter...
- Templating...
- OpenRefine project archive to Google Drive...
- Google Sheets...
- Wikibase edits...
- QuickStatements file
- Wikibase schema
- Pretty Exports
- Stream Exports

- RDF/XML (Pretty)
- Turtle (Pretty)
- TriG (Pretty)
- JSON-LD (Pretty)
- RDF/JSON (Pretty)

Podemos exportar los datos una vez finalizado el proceso de transformación

```
7   @prefix schema: <https://schema.org/> .
8   @prefix vcard:  <http://www.w3.org/2006/vcard/ns#> .
9   @prefix xsd:    <http://www.w3.org/2001/XMLSchema#> .
10
11  <http://example/mapa/3>
12          rdf:type              schema:CreativeWork ;
13          schema:dateCreated    "1579-11-22" ;
14          schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
15          schema:name           "Pintura de Zapotitlan" ;
16          schema:thumbnailURL   "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_32713c87-b23c-474c-ac
17
18  <http://example/mapa/8>
19          rdf:type              schema:CreativeWork ;
20          schema:dateCreated    "1580-04-29" ;
21          schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
22          schema:name           "Pintura de Coatzacualco" ;
23          schema:thumbnailURL   "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_9a78720b-d927-4635-bd
24
25  <http://example/mapa/1>
26          rdf:type              schema:CreativeWork ;
27          schema:dateCreated    "1580-10-26" ;
28          schema:license        "https://creativecommons.org/publicdomain/mark/1.0/" ;
29          schema:name           "Pintura de Cuzcatlan" ;
30          schema:thumbnailURL   "https://curio.lib.utexas.edu/assets/DAMS/utblac/utblac_e38f8f71-f967-46fb-94
```

El fichero RDF en formato TTL se encuentra en la carpeta output del proyecto GitHub

https://github.com/hibernator11/uned-unlocking-workshop/blob/main/output/UNED_metadatos_csv.ttl

Transformación - Código reproducible

GLAM WorkBench

LC Maps for Robots

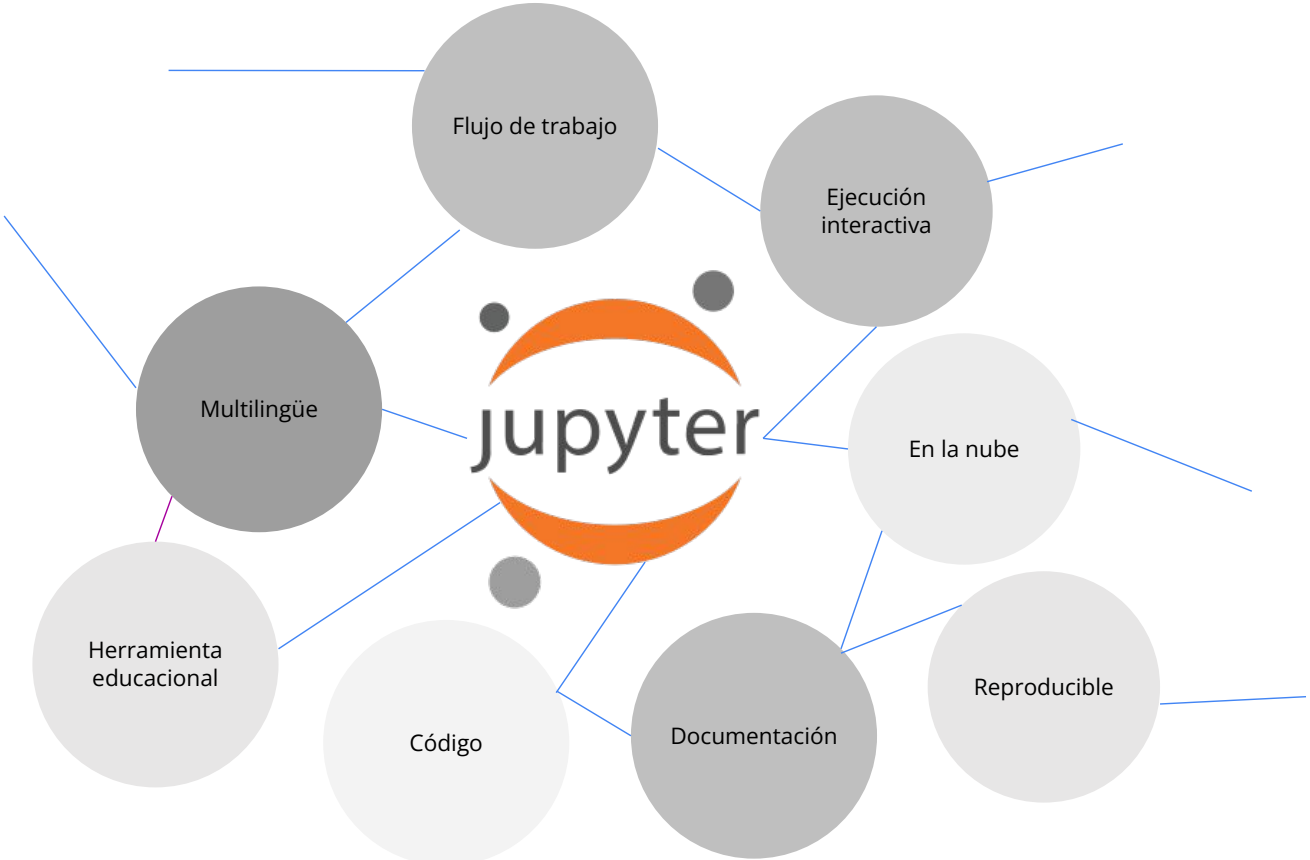
BL Labs

National Library of Scotland

Archives Unleashed Notebooks

LC - Newspapers navigator

Biblioteca Virtual Miguel de Cervantes



dhlab-nova-lisboa

launch binder

This project was originally created for a workshop f … H of Universidade NOVA de Lisboa.

LABORATÓRIO DE HUMANIDADES DIGITAIS

NOVAFCSH

This project provide several examples of reuse based on different technique … Language Processing.

Map representing the nationalities of the authors … de Cervantes

This example is based on the Linked Open Data version of the Biblioteca Vi … Wikidata, enabling the creation of visualisations using the Wikidata SPARQL …

Thanks to OVH, GESIS Notebooks and Curvenote for supporting us!
mybinder.org has updated the base image to Ubuntu 22.04! See the upgrade guide for details.

binder

…rting repository: hibernator11/uned-unlocking … orksho …

1

2

3

https://hub.binder.cu … /hibernator11-un-ocking-workshop-04xnedcb/lab/tree/notebo … cion_UCA

File Edit View Run Kernel Tabs Settings Help

Launcher | Extraccion_UCA2CSV.ipynb

Download GitHub Binder Markdown

Filter files by name

/ notebooks /

Name	Last Modified
Extraccion_...	2 minutes ago
SPARQLAna...	2 minutes ago

Introducción

Este ejemplo se encarga de extraer los metadatos necesarios de la colección. En concreto, vamos a extrae

En primer lugar cargamos las librerías necesarias

[1]:
```
import urllib.request, json
import csv
```

Función que extrae los campos requeridos

[47]:
```
def getCampos(urlmanifest):
    manifest = json.load(urllib.request.urlopen(urlmanifest))

    titulo = ''
    autores = ''
    fecha = ''
    lugar = ''
    coordenadas = ''
    thumbnail = ''
    licencia = ''
```

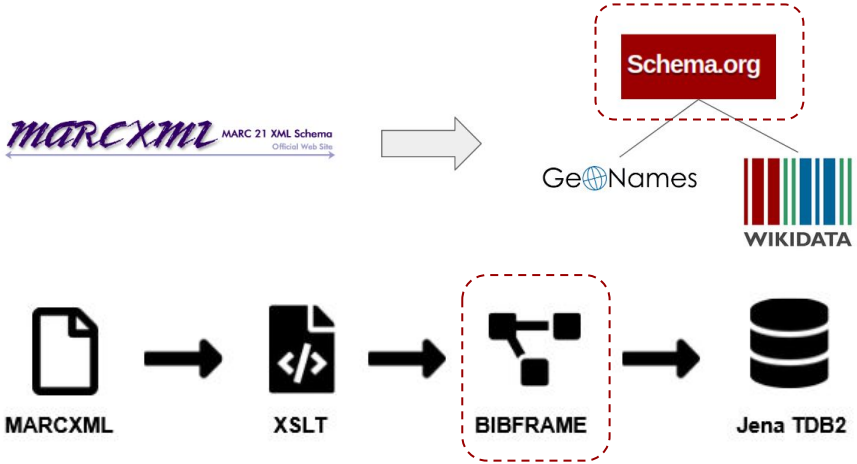
https://github.com/hibernator11/dhlab-nova-lisboa

Transformación

National Librarian's Research Fellowship in Digital Scholarship 2022-23

National Library of Scotland
Leabharlann Nàiseanta na h-Alba

Candela, G. (2023). Towards a semantic approach in GLAM Labs: The case of the Data Foundry at the National Library of Scotland. *Journal of Information Science*. https://doi.org/10.1177/01655515231174386



https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23

Transformación

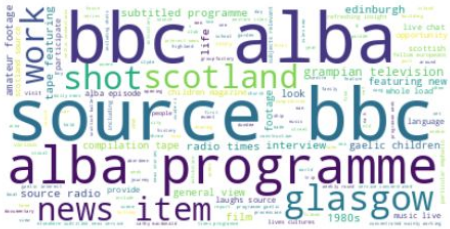
Moving Image Archive

This dataset represents the descriptive metadata from the Moving Image Archive catalogue, which is Scotland's national collection of moving images.

- Data format: metadata available as MARCXML and Dublin Core
- Data source: https://data.nls.uk/data/metadata-collections/moving-image-archive/

The Jupyter Notebooks include a set of examples to reproduce the transformation to RDF and enrichment with external repositories:

- Data extraction
- Exploring the CSV text file
- Transformation to LOD
- Enrichment
- Exploring with SPARQL
- Exploring geographic locations
- Data Quality assessment



Preparation

Import the libraries required to explore the summary of each record included in the dataset to present a word cloud

In [3…
```
from rdflib import Graph, URIRef, Literal, Namespace
from rdflib.namespace import FOAF, RDF, DCTERMS, VOID, DC, SKOS
import pandas as pd
```

Transformation to RDF

*Note: The variable domain could be updated to the domain of the organisation (e.g., https://data.nls.uk).

In [3…
```
domain = 'https://example.org/'
```

First, we instantiate all the namespaces that we will use when defining the RDF data

In [3…
```
g = Graph()
g.bind("foaf", FOAF)
g.bind("rdf", RDF)
g.bind("dcterms", DCTERMS)
g.bind("dc", DC)
g.bind("void", VOID)
g.bind("skos", SKOS)

schema = Namespace("https://schema.org/")
g.bind("schema", schema)

edm = Namespace("http://www.europeana.eu/schemas/edm/")
g.bind("edm", edm)
```

We define the resource National Library of Scotland

In [3…
```
nls = URIRef(domain + "organisation/nls")
g.add((nls, RDF.type, schema.Organization))
g.add((nls, schema.url, URIRef("https://www.nls.uk/")))
g.add((nls, schema.logo, URIRef("https://www.nls.uk/images/nls-logo.png")))
g.add((nls, schema.name, Literal("National Library of Scotland")))
g.add((nls, DC.title, Literal("National Library of Scotland")))
```

https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23

Transformación

National Bibliography of Scotland and BOSLIT

The transformation process is based on the tool marc2bibframe that uses BIBFRAME as main vocabulary to describe the resources.

https://github.com/NLS-Digital-Scholarship/nls-fellowship-2022-23

Loading RDF

The class `JenaTDBLoad` is in charge of loading the RDF into the RDF Jena TDB storage system. The RDF files mu
be places in the folder `rdf`. The following code shows the process:

```
 // Create dataset
Path path = Paths.get(".").toAbsolutePath().normalize();
String dbDir = path.toFile().getAbsolutePath() + "/db/";
Location location = Location.create(dbDir);
Dataset dataset = TDB2Factory.connectDataset(location);

dataset.begin(ReadWrite.WRITE);
Model model = dataset.getDefaultModel();

# update path
Files.walk(Paths.get(path.toFile().getAbsolutePath() +"/rdf"))
        .filter(p -> p.toString().endsWith(".gz"))
        .forEach(p -> {
            logger.info(p.toFile().getAbsolutePath());
            try {
                RDFDataMgr.read(model, p.toFile().getAbsolutePath(), Lang.RDFXML);
            }catch (Exception e){
                logger.error(p.toFile().getAbsolutePath() + e.getMessage());
            }
        });
dataset.commit();

// Releasing dataset resources
dataset.close();
```

https://github.com/hibernator11/nls-jena-tdb

Java

Maven™

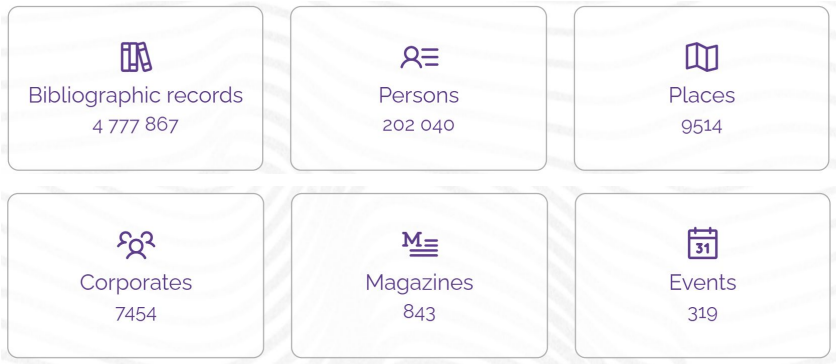
Apache Jena

Transformación

ATRIUM

https://atrium-research.eu/

Facilitating access to digital research infrastructures and advancing frontier knowledge in the arts and humanities — across disciplines, languages and media.



literarybibliography.eu

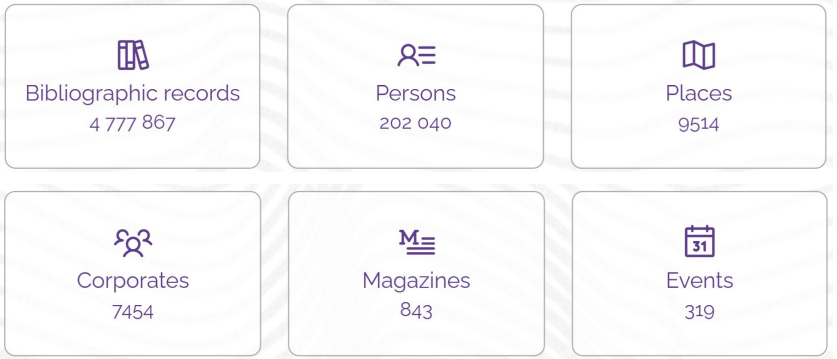
Transformación

DH2025
ACCESSIBILITY & CITIZENSHIP

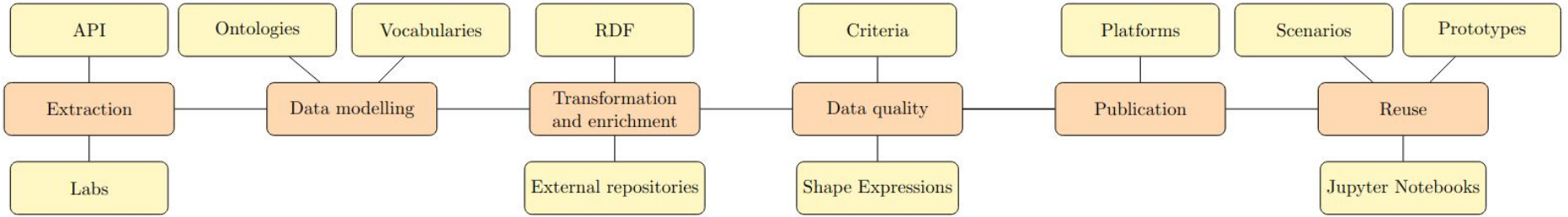
ATRIUM

https://atrium-research.eu/

Facilitating access to digital research infrastructures and advancing frontier knowledge in the arts and humanities — across disciplines, languages and media.



literarybibliography.eu



ATRIUM project

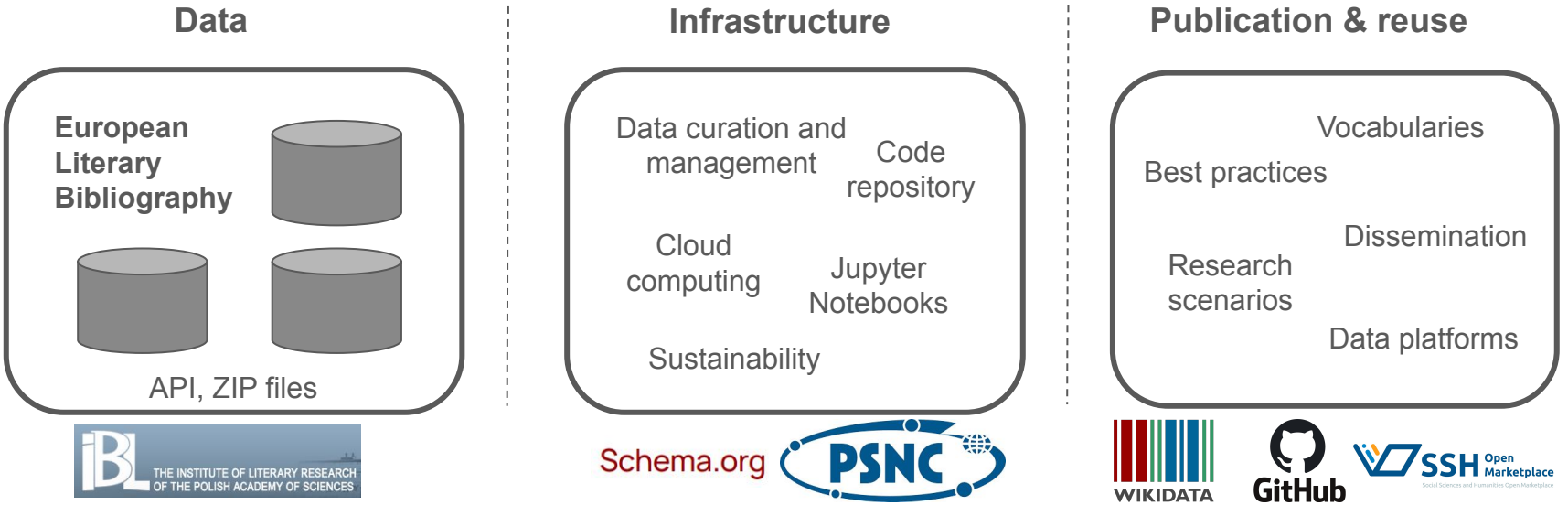
Candela, G., Rosiński, C., & Margraf, A. (2024). A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography.
https://doi.org/10.5281/zenodo.14106707

A reproducible approach

The **Jupyter Notebooks** include a set of examples to reproduce the transformation to RDF.

- Data extraction: extraction of the data from the biblio.json file to create a CSV file.
- Data exploration: exploration of the original biblio.json file.
- Transformation to LOD: transformation of the original data using the CSV file to RDF using as main vocabulary Schema.org.
- Data exploration using SPARQL: exploration using SPARQL of the RDF data generated.
- Data quality ShEx: validation of the RDF data generated using Shape Expressions. In order to check how to test the ShEx against the RDF generated you can see the Data quality assess notebook.

Transformación



Architecture defined to apply the framework proposed in this work. It focuses on three main elements: data, infrastructure, and publication

Transformación

Dataset	Classes	Properties	No. triples	No. external links
Adam Mickiewicz	7	19	47,631	863
Miguel de Cervantes	5	19	44,389	749
National Library of Spain	7	19	10,515,016	67,563

Scenario 1: Comparative analysis of provincial Spanish novels featuring vampires
Scenario 2: Republican writers who migrated during the Spanish Civil War
Scenario 3: Geographical distribution of publications dedicated to particular Spanish writers

Transformations: A DARIAH Journal

Article

A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography

Gustavo Candela (1), Cezary Rosiński (2), Arkadiusz Margraf (3)

(1) University of Alicante
(2) Institute of Literary Research of the Polish Academy of Sciences
(3) Institute of Bioorganic Chemistry of the Polish Academy of Sciences

Abstract

EN GLAM (Galleries, Libraries, Archives and Museums) institutions host rich content that is provided in the form of digital collections. Bibliographic databases are collections of references focused on a particular topic that can be used to apply Digital Humanities (DH) methods. Recent approaches such as Collections as Data and Labs promote the publication of digital collections supporting computational use.This work aims to provide a framework for publishing and reusing digital collections based on literary bibliographies published by GLAM institutions in order to make them suitable for computational use in the form of Collections as Data, in particular, in the context of the European Literary Bibliography. It also describes the infrastructure used and DH research scenarios to illustrate how the results can be reused for different goals. Digital curators and DH researchers interested in making their datasets available in the form of Collections as Data are the intended audience of this work.

Download article Open on Zenodo
Cite Metadata Share

Publication details

Published on	June 16, 2025
Accepted on	April 3, 2025
Submitted on	November 12, 2024

Volume 1
Workflows

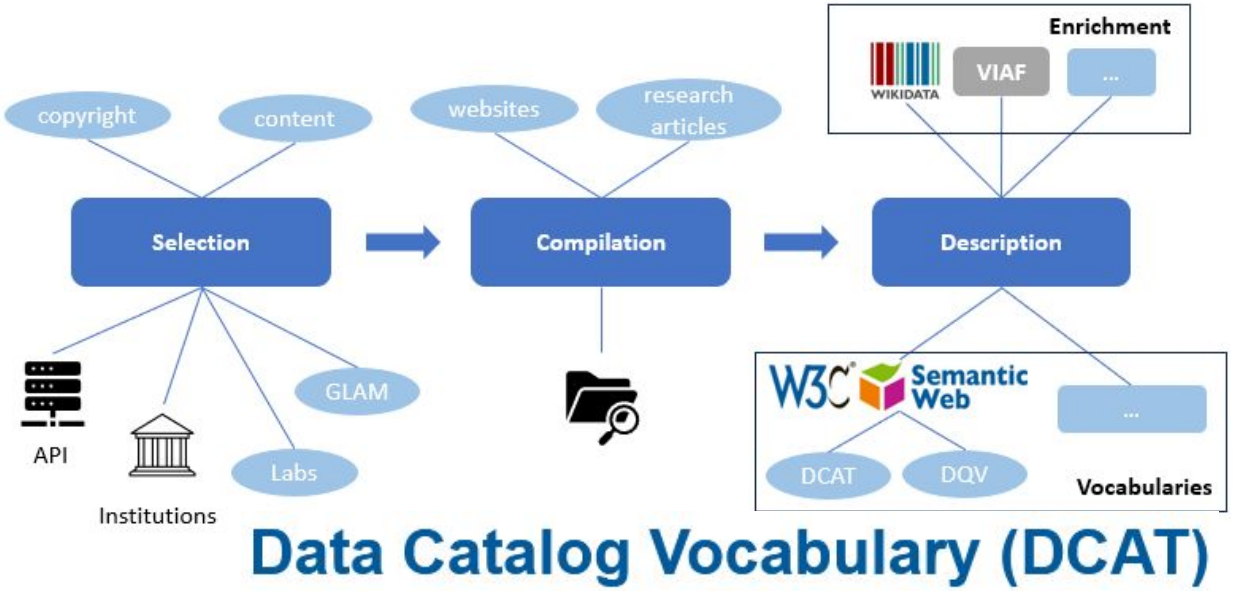
DOI
10.46298/transformations.14729

https://doi.org/10.46298/transformations.14729

https://marketplace.sshopencloud.eu/workflow/I3JvP6

Transformación

Espacios de datos de patrimonio cultural



Data Catalog Vocabulary (DCAT)

Workflow steps (10)

1	Provide a clear license and terms of use allowing reuse of the dataset without restrictions	Expand
2	Provide a suggested citation for the dataset so reusers are aware of how to cite it	Expand
	about the dataset	Expand
	make available the dataset for the public	Expand
	o demonstrate how the dataset can be reused	Expand
	for the dataset for a better understanding of how to reuse	Expand
	le metadata about the content provided in the dataset	Expand
	rative-edition platform to include the information about the	Expand
	means of an existing API	Expand
	sent and describe the dataset to encourage its reuse	Expand

https://github.com/hibernator11/dcat-glam-catalog

Transformación

Spanish-Civil-War-KGs

Extracting Knowledge Graphs as Collections as Data using the Spanish Civil War as case study

DOI 10.5281/zenodo.16747522

Introduction

This project intends to analyse the options and potential of Wikidata to extract Knowledge Graphs as Collections as Data using the Spanish Civil War as case study.

The data was retrieved from the National Archives in Spain, and enriched in Wikidata.

The authors of this work are:

- Gustavo Candela, University of Alicante, Spain
- Meltem Dişli, Hacettepe University, Türkiye
- Paul Spence, King's College London, United Kingdom
- Cezary Rosiński, Institute of Literary Research of Polish Academy of Sciences, Poland
- Arkadiusz Margraf, Poznan Supercomputing and Networking Center affiliated to Institute of Bioorganic Chemistry of the Polish Academy of Sciences, Poland

We provide 4 examples of use covering different topics and content:

- Artists from the period of the Spanish Republic and Spanish Civil War
- Refugee ships on which Spanish exiles embarked during the Spanish Exile and the Spanish Civil War
- French refugee camps, which concentrated the Spanish Republican exiles of 1939
- Retrieving bibliographic records related to the Spanish Civil War from the Biblioteca Virtual Miguel de Cervantes LOD repository

Casos de uso basados en la extracción de datos y la Guerra Civil Española

Transformación

¿Cómo buscar en varios repositorios y evitar silos de datos?

Creating Collections as Data Using Federated Queries

This repository provides a series of **Jupyter notebooks** designed for performing **federated queries** across **Wikidata** and other **linked open data (LOD)** endpoints, with a focus on querying bibliographic and author-related data.

These notebooks support our paper *From Linked Open Data to Collections as Data: A Reproducible Framework Using Federated Queries* and aim to present reproducible examples of query design strategies other than structured real-world use in library and information studies, digital humanities and other domains.

launch binder

Materials in this Repository

The repository is organized into the following sections:

- **Query Building**:
  - *4 examples* showcasing a query-building workflow for three different endpoints:
    - Biblioteca Nacional de España (BNE) Query Building
    - Bibliothèque nationale de France (BNF) Query Building
    - Biblioteca Virtual Miguel de Cervantes (BVMC) Query Building
    - Wikidata_map_viz, an exploratory query visualizing authors' places of birth from the Spanish Golde Age (wd:Q530936).
- **Single Author, Multiple Works**:
  - *2 examples* focusing on retrieving multiple works by a single author:
    - Jorge Juan y Santacilia
    - Maria Zayas y Sotomayor
- **Multiple Authors, Multiple Works**:
  - *1 example* demonstrating queries involving multiple authors and their works,
    - Spanish Golden Age

Ejemplos de sentencias SPARQL federadas

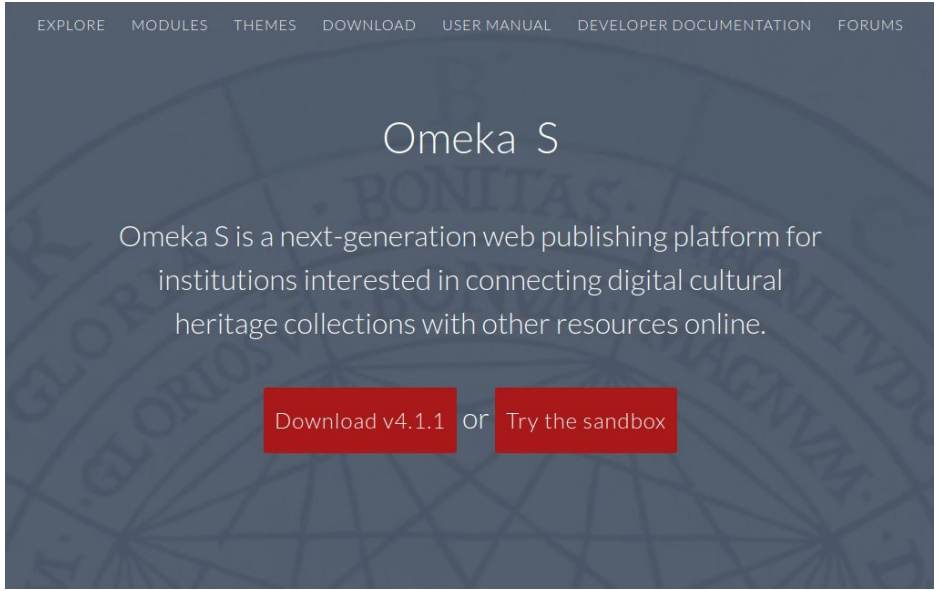
Transformación

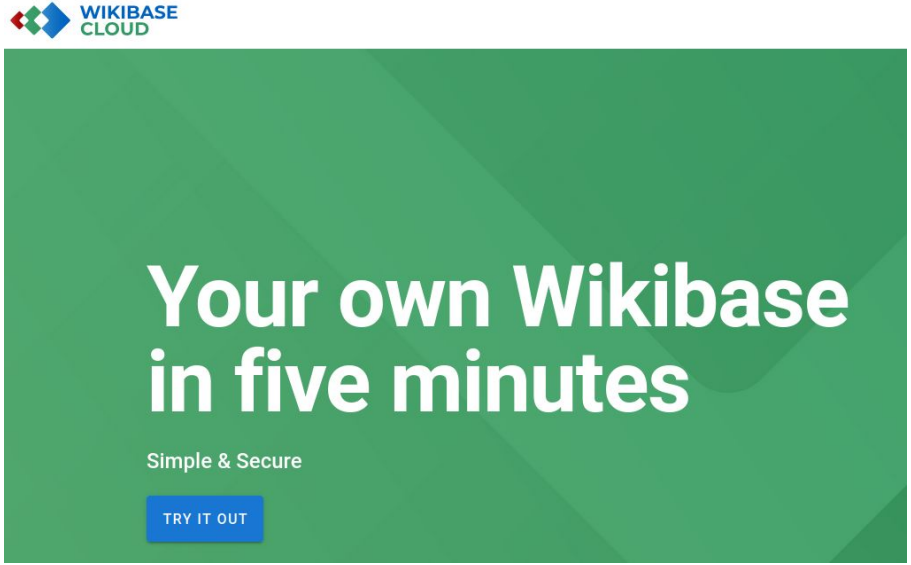
¿Cómo buscar en varios repositorios y evitar silos de datos?

```
PREFIX bne-def: <https://datos.bne.es/def/>
PREFIX dcterms: <http://purl.org/dc/terms/>

SELECT ?author ?authorLabel ?work ?workLabel ?edition ?placeOfProduction ?yearOfPublication ?langCode
WHERE {
   ?author wdt:P950 ?id .
   ?author wdt:P135 wd:Q530936 .
   ?author rdfs:label ?authorLabel.  FILTER(LANG(?authorLabel) = "en").
   BIND(uri(concat("https://datos.bne.es/resource/", ?id)) as ?bneID)
   SERVICE <http://datos.bne.es/sparql> {
      ?bneID bne-def:OP5001 ?work .
      ?work rdfs:label ?workLabel .
      OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition bne-def:P3003 ?placeOfProduction}
      OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition bne-def:P3006 ?yearOfPublication}
      OPTIONAL {?work bne-def:OP1002 ?m . ?m bne-def:OP2001 ?edition . ?edition dcterms:language ?langCode}
   }
}
LIMIT 1000
```

Obras de la BNE de autores pertenecientes al Siglo de Oro en Wikidata

Transformación





https://omeka.org/s/

https://www.wikibase.cloud/

Transformación

DHWIKI

Buscar en DHWiki

Main Page

Portada | Discusión | Ver código fuente | Historial

Welcome do **DHwiki Wikibase**. This mediaWiki installation, hosted on Wikibase Cloud, includes a Wikibase instance, and is maintained since September 2024 by the **DARIAH-EU WG DHwiki**, in the framework of the DARIAH ERIC.

Please read the public announce of the WG creation. You find a short set of slides about our group here: https://doi.org/10.5281/zenodo.16041468 You find dates and agendas of our group meetings here: Group meetings.

This Wikibase contains:

- Documentation pages about Wikibase in a Digital Humanities context (in the making.)
- a LOD version of our group bibliography (a Zotero collection available here (see list of contributors));
- information about group members;
- textual descriptions of Wikibase instances related to the digital humanities (cf. Category:Wikibase instances)
- resources related to Wikibase and its uses in the digital humanities (cf. Resource directory).

The group has set up a **Zenodo community**, where you find our papers and presentation slides: https://zenodo.org/communities/dariah_dhwiki/

We also are having experiments with a set of Jupyter notebooks.

Sumario ocultar

1 What's in DHwiki Wikibase?
1.1 Bibliographical items
1.1.1 Bibliographical items domains
1.2 Persons
1.2.1 Authors
1.2.2 Group members
1.2.2.1 Group members image gallery
1.3 Wikibase instances described on Wikibase World and on DHwiki Wikibase
1.4 Classes and instances on this Wikibase

Navegación

Página principal
Cambios recientes
Página aleatoria
Ayuda sobre MediaWiki

Wikibase

New Item
New Property
New Schema
All Properties
Query Service
Cradle
QuickStatements

Herramientas wiki

Páginas especiales

Findable Accessible Interoperable Reusable

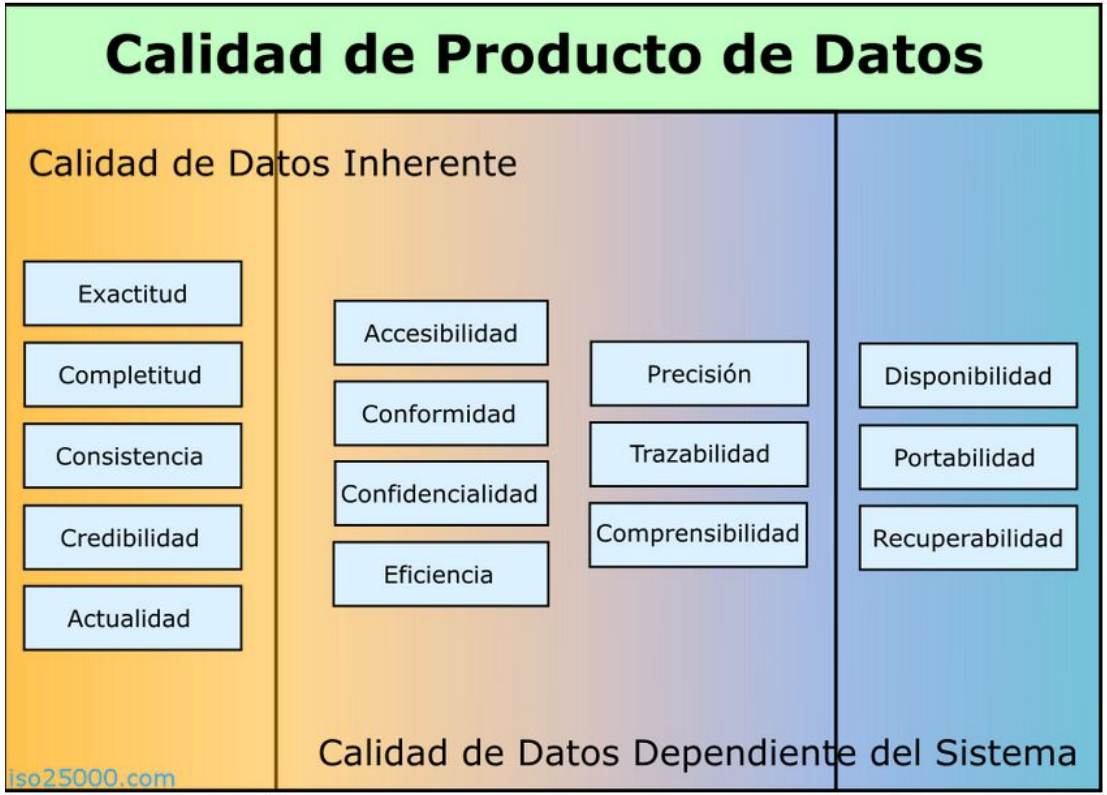
SPARQL

WIKIDATA

jupyter

https://dhwiki.wikibase.cloud/wiki/Main_Page

Calidad de datos



https://www.iso25000.com/index.php/normas-iso-25000/iso-25012

¿Qué medimos?

Calidad de datos

Ejemplos

https://shex.io/

A Shape Expression approach for assessing the quality of Linked Open Data in libraries

https://doi.org/10.3233/SW-210441

A reproducible framework to assess LOD

https://github.com/hibernator11/lod-quality-reproducible

Evaluating the quality of linked open data in digital libraries

https://doi.org/10.1177/0165551520930951

An automatic data quality approach to assess semantic data from cultural heritage institutions

https://doi.org/10.1002/asi.24761

A Reproducible Framework to Assess LOD Quality

Assessment results

Running in the cloud

Create visual interface

```
children = []
titles = []
tab = widgets.Tab()
children.insert(0,introductionLayout)
children.insert(1,criteriaLayout)
children.insert(2,assessmentResultsLayout)
tab.children = children

titles.insert(0, 'Repository')
titles.insert(1, 'Data quality criteria')
titles.insert(2, 'Assessment results')
tab.titles = titles

tab
```

4

Repository | Data quality criteria | Assessment results

SPARQL: https://data.bnf.fr/sparql

Title: National Library of France

Description: Reference information about authors, work

5

Repository | Data quality criteria | Assessment results

6

This section shows the data quality assessment results obtained for the LOD repositories selected. The results are retrieved from the ttl files in the data folder using the properties dqv:value.

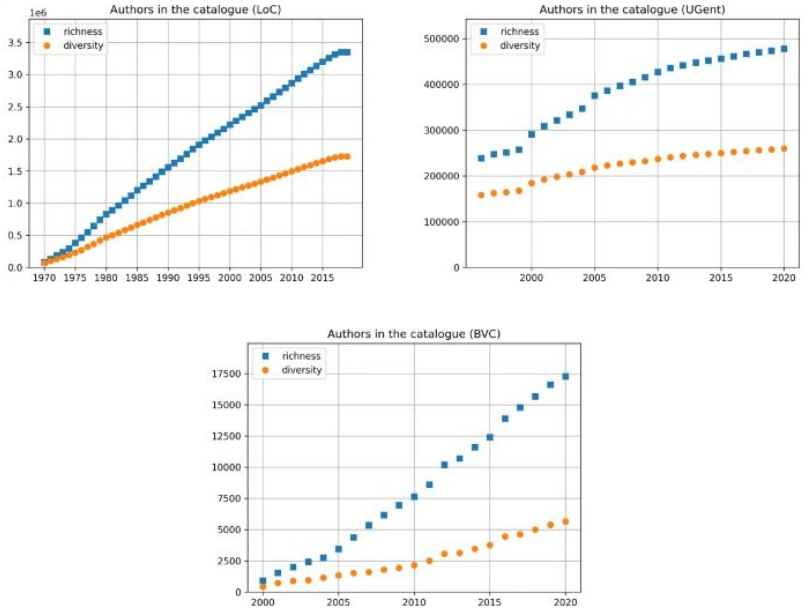
Dimension	Criterion	Result
Availability	Availability	correct
Licensing	Licensing	not satisfied
Interlinking	Interlinking	correct
Security	Security	correct

https://github.com/hibernator11/lod-quality-reproducible

Calidad

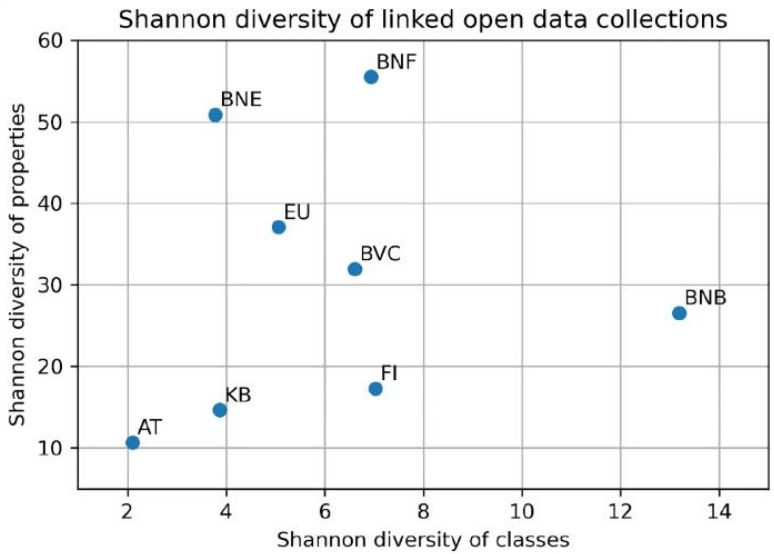
C. Carrasco, R., Candela, G. & Marco-Such, M. Measuring the diversity of data and metadata in digital libraries. *Int J Digit Libr* **26**, 5 (2025). https://doi.org/10.1007/s00799-025-00411-1

Fig. 8



Cumulative number of authors and Shannon diversity of the authors in the catalogue as a function of the year the MARC record was included in the catalogue extracted from the MARC field 008

Fig. 10



Shannon diversity of classes and properties in *Linked Open Data* published by libraries

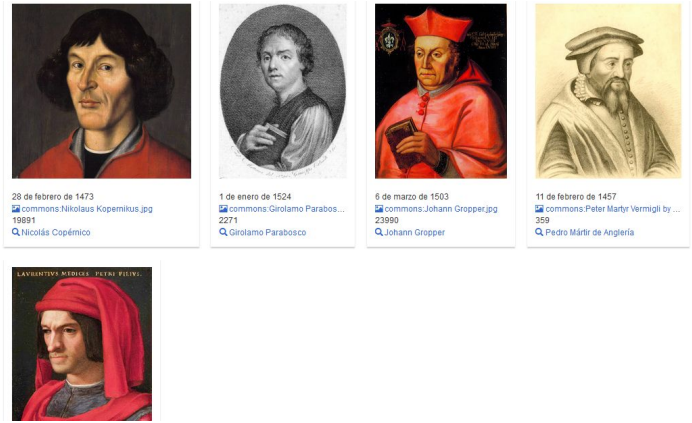
Visualización

Visualizaciones

- #defaultView:ImageGrid

https://w.wiki/8922

```
#defaultView:ImageGrid
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento .
  ?autor wdt:P18 ?imagen

  FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
         && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .

  SERVICE wikibase:label {
          bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es". }
}
LIMIT 10
```

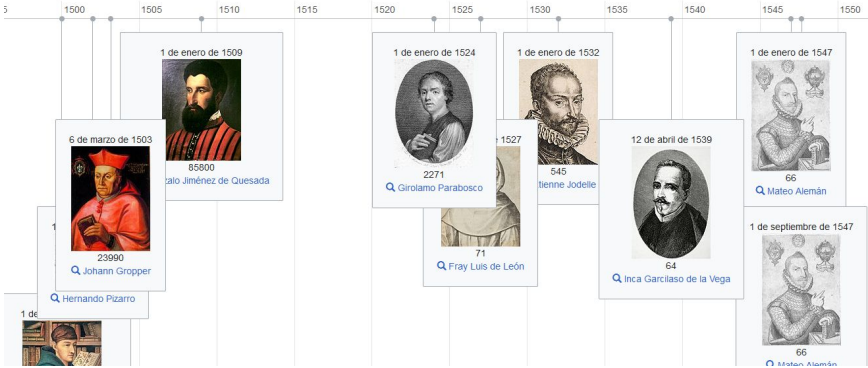


Visualización

Visualizaciones

- #defaultView:Timeline

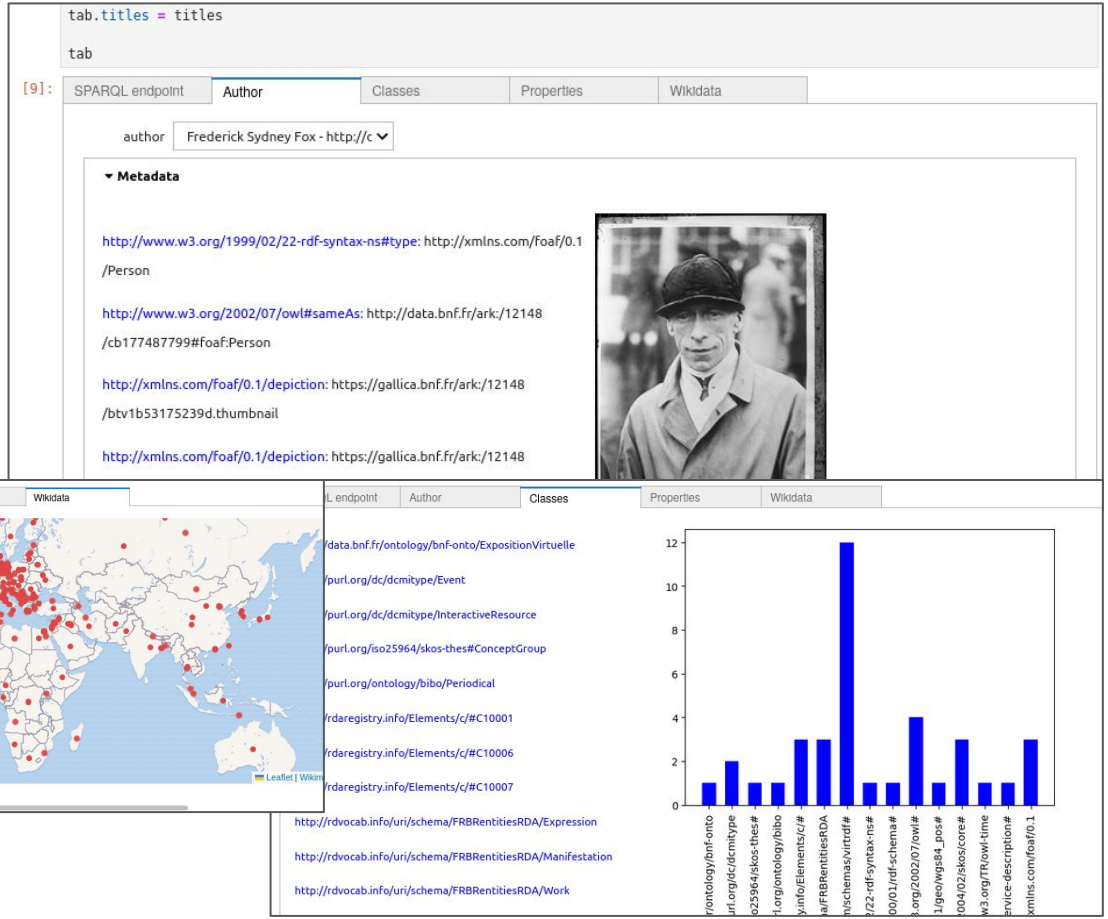
https://w.wiki/88$y




```
#defaultView:Timeline
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento .
  ?autor wdt:P18 ?imagen

  FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
         && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .

  SERVICE wikibase:label {
          bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es". }
}
LIMIT 10
```

Visualización

A **reproducible approach** to **browse** Linked Open Data in Cultural Heritage



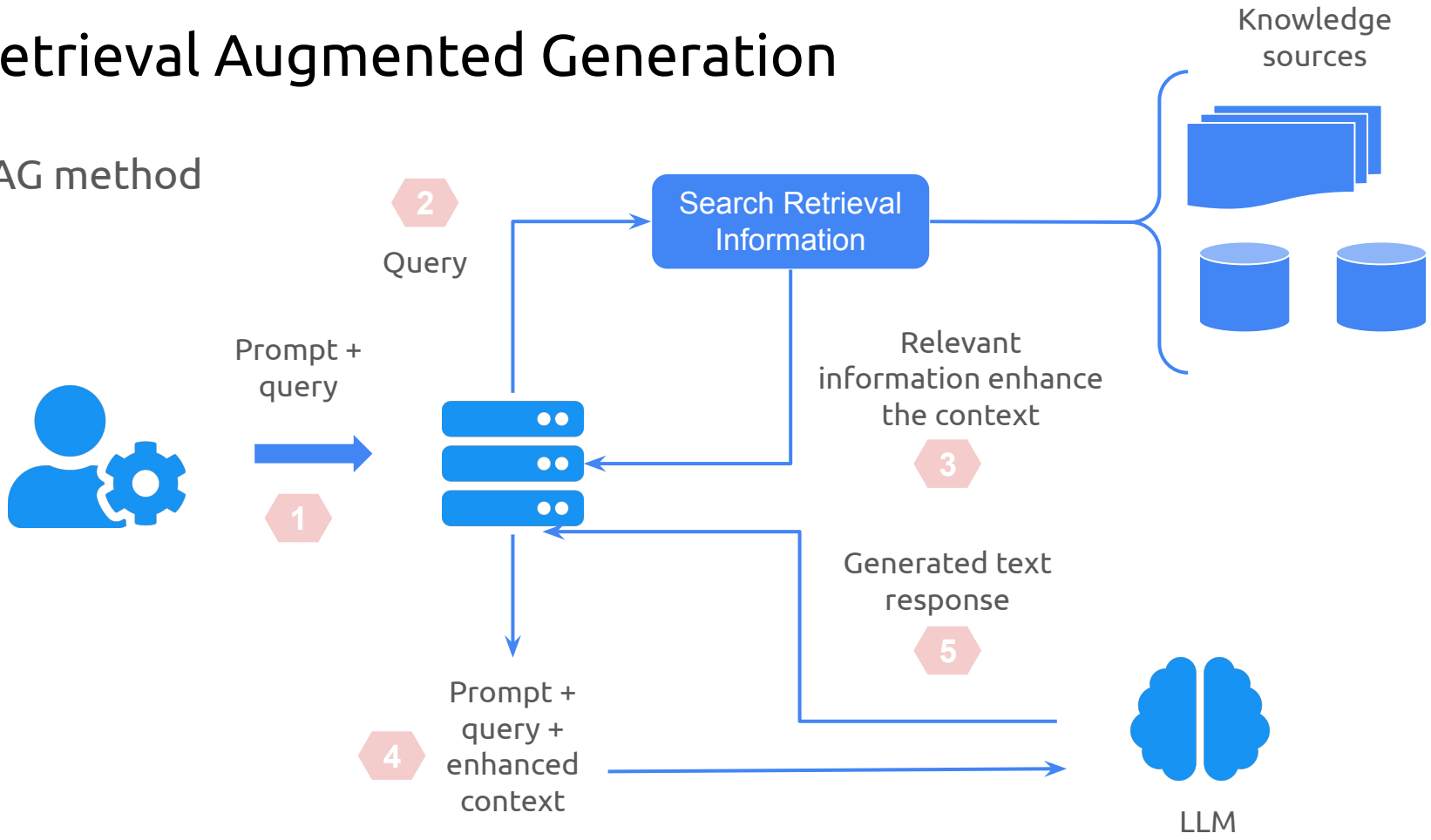
Schema.org

https://github.com/hibernator11/visual-ch-lod

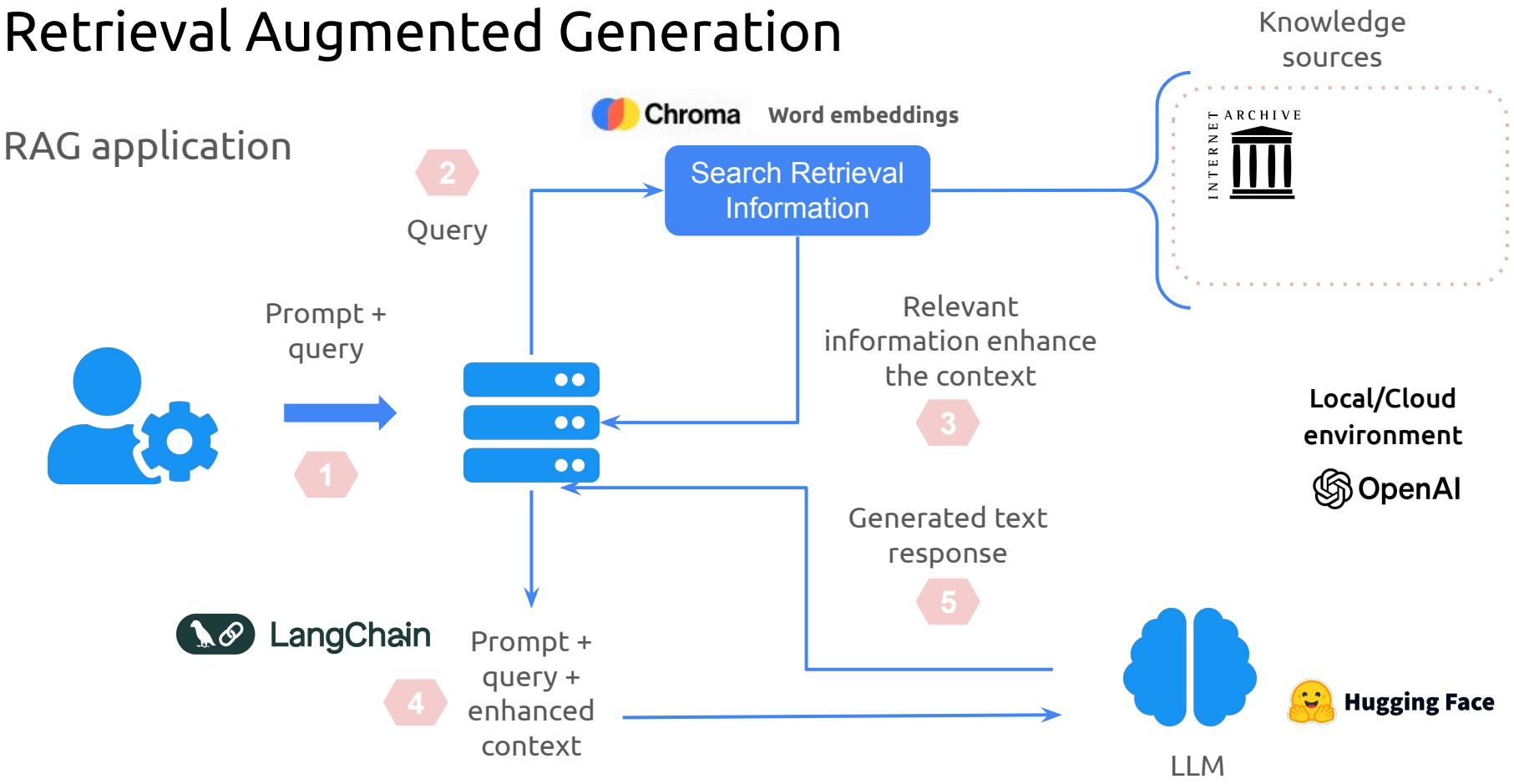
https://doi.org/10.1145/3707647

Retrieval Augmented Generation

RAG method



Retrieval Augmented Generation



Visualización y reutilización de datos

Adquisición y preparación de datos

Universitat d'Alacant
Universidad de Alicante

Contenido

- * Ejemplos de transformación
- Introducción
- Reutilización
  - Visión por computador
  - Calidad de datos
- Visualización
  - Herramientas
  - Ejemplos

Transformación

Moving Image Archive

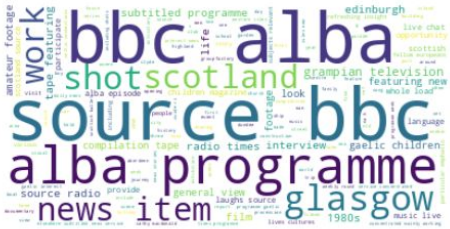
This dataset represents the descriptive metadata from the Moving Image Archive catalogue, which is Scotland's national collection of moving images.

- Data format: metadata available as MARCXML and Dublin Core
- Data source: https://data.nls.uk/data/metadata-collections/moving-image-archive/

The Jupyter Notebooks include a set of examples to reproduce the transformation to RDF and enrichment with external repositories:

- Data extraction
- Exploring the CSV text file
- Transformation to LOD
- Enrichment
- Exploring with SPARQL
- Exploring geographic locations
- Data Quality assessment

Lee CSV y genera un fichero TTL con las tripletas



Preparation

Import the libraries required to explore the summary of each record included in the dataset to present a word cloud

In [3…
```
from rdflib import Graph, URIRef, Literal, Namespace
from rdflib.namespace import FOAF, RDF, DCTERMS, VOID, DC, SKOS
import pandas as pd
```

Transformation to RDF

*Note: The variable domain could be updated to the domain of the organisation (e.g., https://data.nls.uk).

In [3…
```
domain = 'https://example.org/'
```

First, we instantiate all the namespaces that we will use when defining the RDF data

In [3…
```
g = Graph()
g.bind("foaf", FOAF)
g.bind("rdf", RDF)
g.bind("dcterms", DCTERMS)
g.bind("dc", DC)
g.bind("void", VOID)
g.bind("skos", SKOS)

schema = Namespace("https://schema.org/")
g.bind("schema", schema)

edm = Namespace("http://www.europeana.eu/schemas/edm/")
g.bind("edm", edm)
```

We define the resource National Library of Scotland

In [3…
```
nls = URIRef(domain + "organisation/nls")
g.add((nls, RDF.type, schema.Organization))
g.add((nls, schema.url, URIRef("https://www.nls.uk/")))
g.add((nls, schema.logo, URIRef("https://www.nls.uk/images/nls-logo.png")))
g.add((nls, schema.name, Literal("National Library of Scotland")))
g.add((nls, DC.title, Literal("National Library of Scotland")))
```

https://github.com/hibernator11/nls-fellowship-2022-23

ATRIUM project

Candela, G., Rosiński, C., & Margraf, A. (2024). A reproducible framework to publish and reuse Collections as data: the case of the European Literary Bibliography.
https://doi.org/10.5281/zenodo.14106707

A reproducible approach

The **Jupyter Notebooks** include a set of examples to reproduce the transformation to RDF.

- Data extraction: extraction of the data from the biblio.json file to create a CSV file.
- Data exploration: exploration of the original biblio.json file.
- Transformation to LOD: transformation of the original data using the CSV file to RDF using as main vocabulary Schema.org.
- Data exploration using SPARQL: exploration using SPARQL of the RDF data generated.
- Data quality ShEx: validation of the RDF data generated using Shape Expressions. In order to check how to test the ShEx against the RDF generated you can see the Data quality assess notebook.

ATRIUM

jupyter

THE INSTITUTE OF LITERARY RESEARCH OF THE POLISH ACADEMY OF SCIENCES

PSNC

Reutilización

Problema general…

¿Qué **datos** tenemos? ¿**Formato**? ¿Enriquecidos?

¿Es necesario realizar una **limpieza** de los datos?

¿Qué quiero **analizar**? ¿Cuál es el **objetivo**?

¿Quién es mi **audiencia**? ¿Estudiantes? ¿Investigadores?

¿Qué **infraestructura** tengo para desplegar mi código? ¿Hay **límites** de memoria o espacio?

¿Qué **licencia** tienen los datos?

OPEN A GLAM LAB

https://www.glamlabs.io/

Visión por computador

Problemas comunes

Object detection is the task of detecting instances of objects of a specific class within an image or video.

Classification

Detection

Segmentation

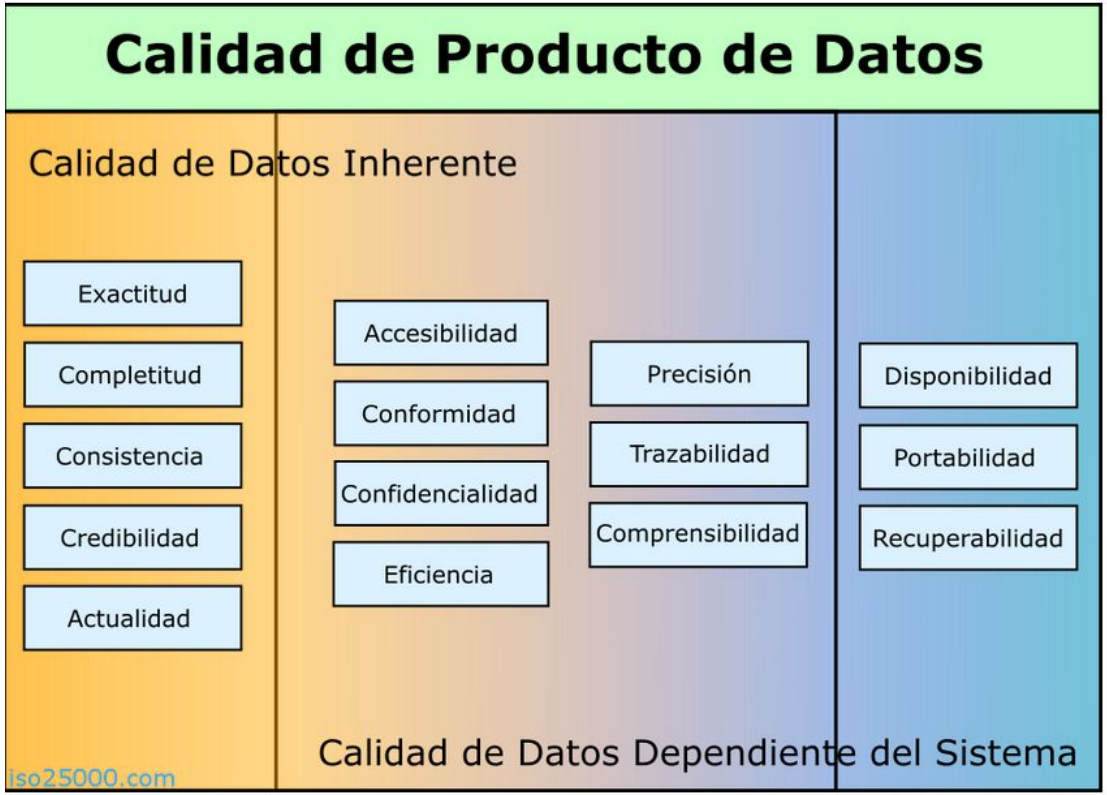
https://www.freecodecamp.org/news/how-to-detect-objects-in-images-using-yolov8/

Visión por computador

https://yolov8.com/

https://github.com/hibernator11/workshop-notebooks-dhnb2024

Calidad de datos



https://www.iso25000.com/index.php/normas-iso-25000/iso-25012

¿Qué medimos?

Calidad de datos

Ejemplos

https://shex.io/

A Shape Expression approach for assessing the quality of Linked Open Data in libraries

https://doi.org/10.3233/SW-210441

A reproducible framework to assess LOD

https://github.com/hibernator11/lod-quality-reproducible

Evaluating the quality of linked open data in digital libraries

https://doi.org/10.1177/0165551520930951

An automatic data quality approach to assess semantic data from cultural heritage institutions

https://doi.org/10.1002/asi.24761

Calidad de datos

1 **Data quality criteria**

Data quality dimensions and criteria based on previous work

2 **Data modelling**

Describing the results using ontologies promoted by the community such as the Data quality Vocabulary

3 **Assessment**

Reproducible code and assessment results

jupyter

Calidad de datos

Running in the cloud

Create visual interface

```
children = []
titles = []
tab = widgets.Tab()
children.insert(0,introductionLayout)
children.insert(1,criteriaLayout)
children.insert(2,assessmentResultsLayout)
tab.children = children

titles.insert(0, 'Repository')
titles.insert(1, 'Data quality criteria')
titles.insert(2, 'Assessment results')
tab.titles = titles

tab
```

4

Repository | Data quality criteria | Assessment results

SPARQL: https://data.bnf.fr/sparql

Title: National Library of France

Description: Reference information about authors, work

5

Repository | Data quality criteria | Assessment results

6

This section shows the data quality assessment results obtained for the LOD repositories selected. The results are retrieved from the ttl files in the data folder using the properties dqv:value.

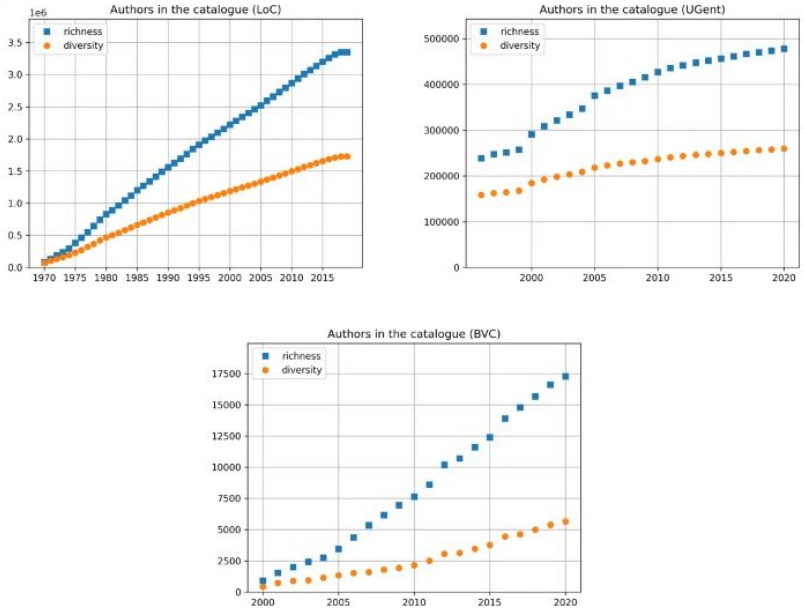
Dimension	Criterion	Result
Availability	Availability	correct
Licensing	Licensing	not satisfied
Interlinking	Interlinking	correct
Security	Security	correct

https://github.com/hibernator11/lod-quality-reproducible

Calidad de datos

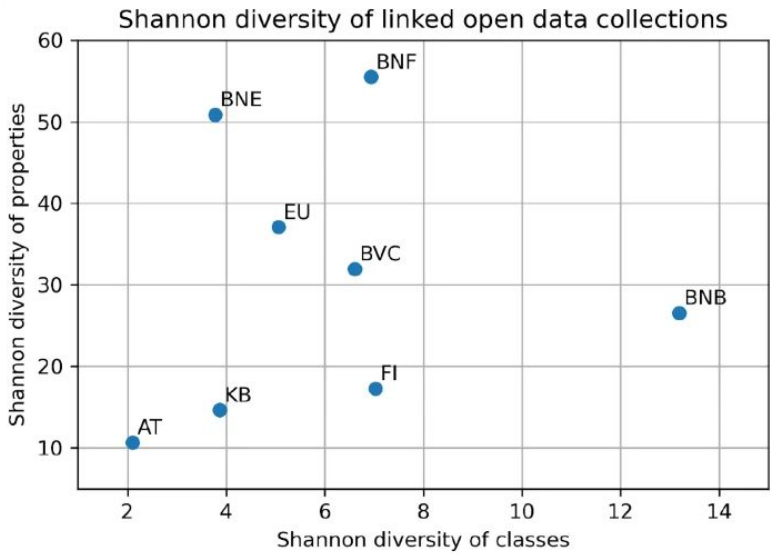
C. Carrasco, R., Candela, G. & Marco-Such, M. Measuring the diversity of data and metadata in digital libraries. *Int J Digit Libr* **26**, 5 (2025). https://doi.org/10.1007/s00799-025-00411-1

Fig. 8



Cumulative number of authors and Shannon diversity of the authors in the catalogue as a function of the year the MARC record was included in the catalogue extracted from the MARC field 008

Fig. 10



Shannon diversity of classes and properties in *Linked Open Data* published by libraries

Visualización

La generación de datos crece de forma exponencial

Los datos deben ser limpiados, filtrados, organizados, resumidos y analizados para facilitar la toma de decisiones

El enriquecimiento con repositorios externos puede facilitar la inclusión de información adicional

Visualización

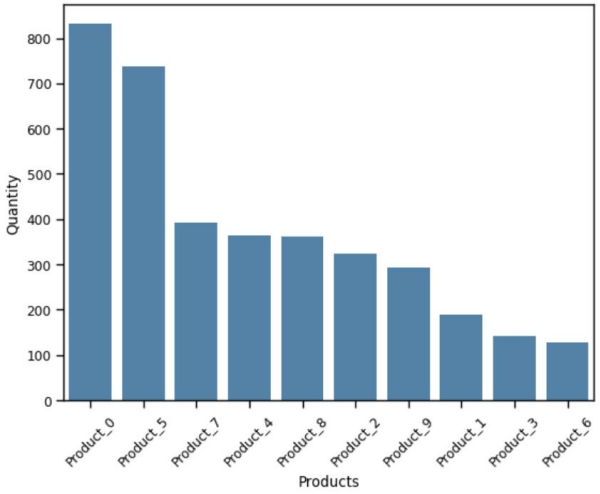
Las visualizaciones dependen de los objetivos perseguidos. Algunos de los más comunes son:

- Ranking
- Comparativas
- Correlaciones
- Distribución
- Evolución en el tiempo

Visualización

Ranking

- Trata de mostrar la relación de **orden** en un conjunto de datos

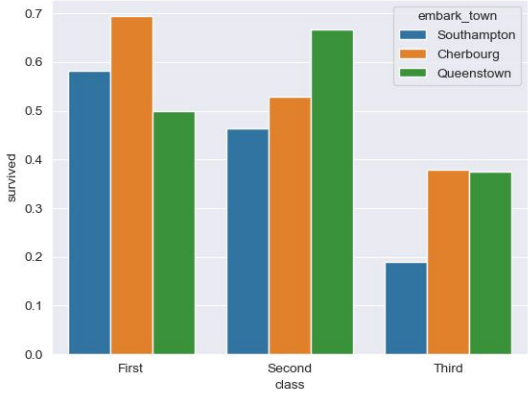


Bar Chart

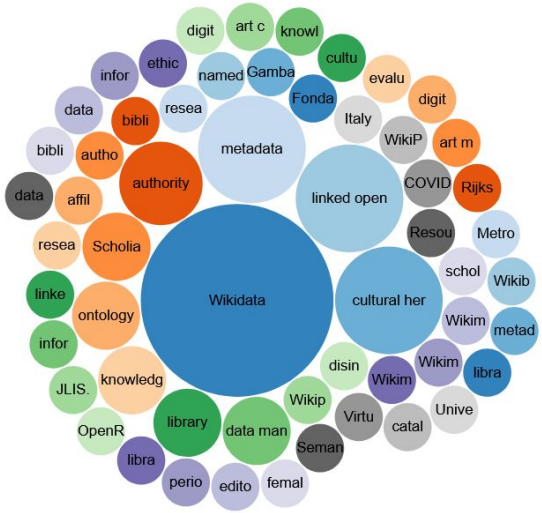
Visualización

Comparativa

- Enfrenta valores numéricos asociados a distintos elementos individuales o grupos



Grouped Bar Chart

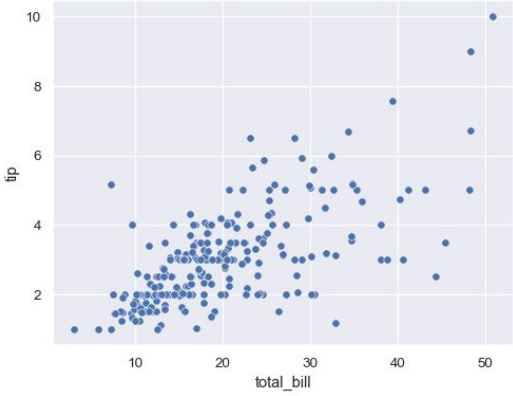


Bubble Chart

Visualización

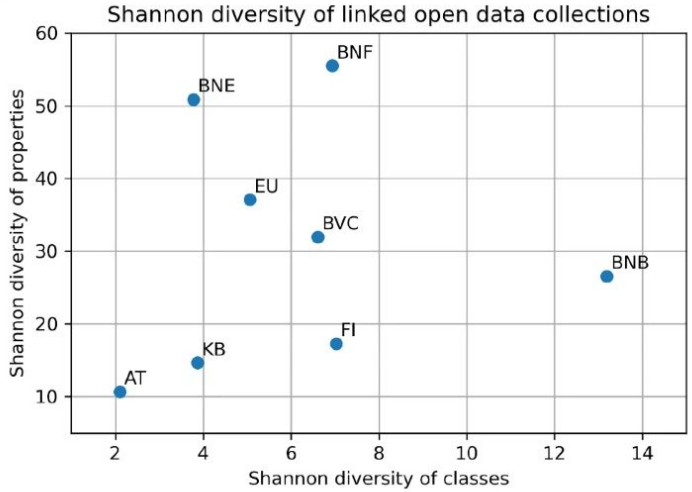
Correlación

- Comprueba la relación entre pares de variables



Scatter plot

Fig. 10



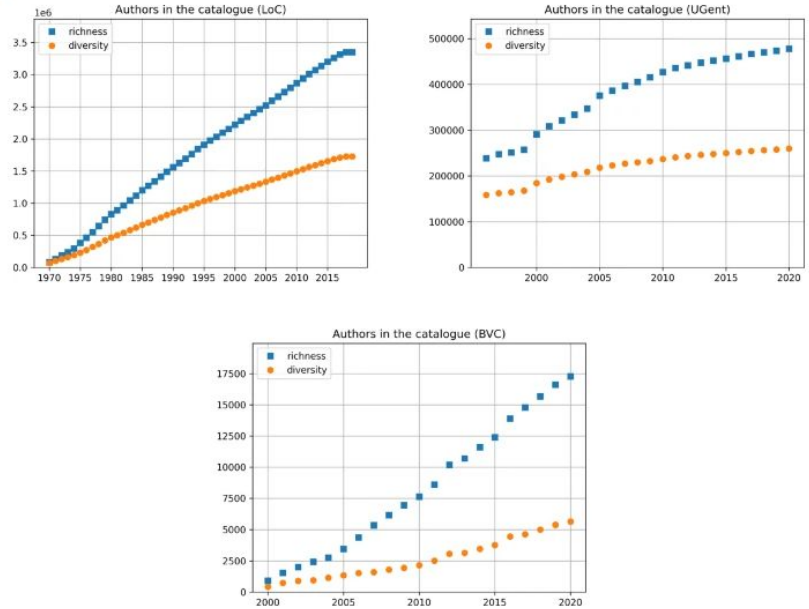
Shannon diversity of classes and properties in *Linked Open Data* published by libraries

Visualización

Evolución en el tiempo

- Reflejan el cambio de valor de una variable con respecto al tiempo

Fig. 8



Cumulative number of authors and Shannon diversity of the authors in the catalogue as a function of the year the MARC record was included in the catalogue extracted from the MARC field 008

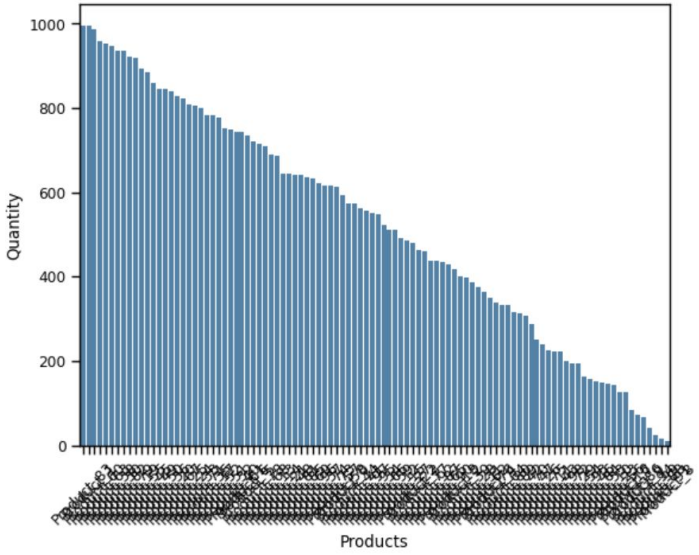
Visualización

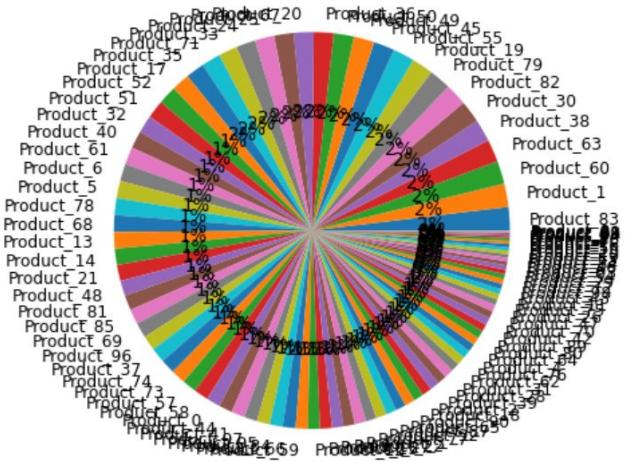
Selección de visualizaciones

- ¿Cómo seleccionar una visualización adecuada?
  - Finalidad:
    - ¿Análisis a lo largo del tiempo? ¿Distribución? ¿Comparación de valores?
  - Público objetivo:
    - ¿Experto en visualizaciones? ¿Experto en dominio?
  - Uso:
    - Exploratorio / Comunicativo
  - Características del dataset:
    - ¿Qué tipo de datos se van a utilizar? ¿Qué volumen de datos se está manejando? ¿Cuántos valores distintos hay? ¿Cuántas dimensiones entran en juego?

Visualización

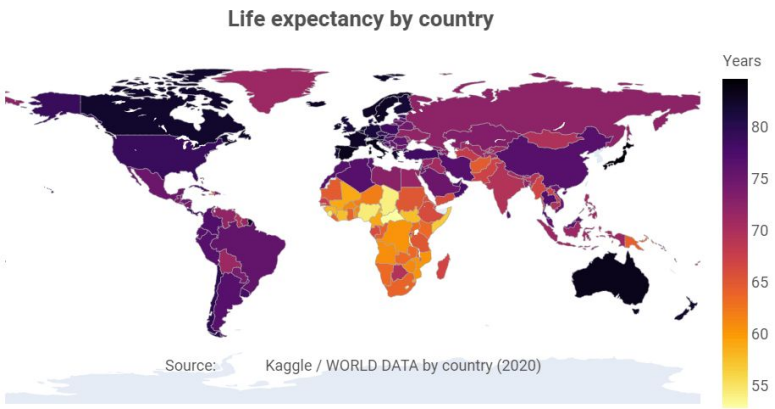
Problema **volumen** de datos

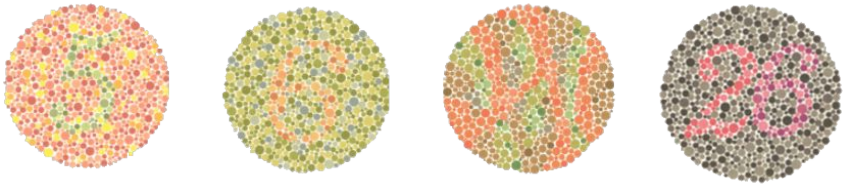




Visualización

Problema **uso del color, daltonismo, choques culturales**



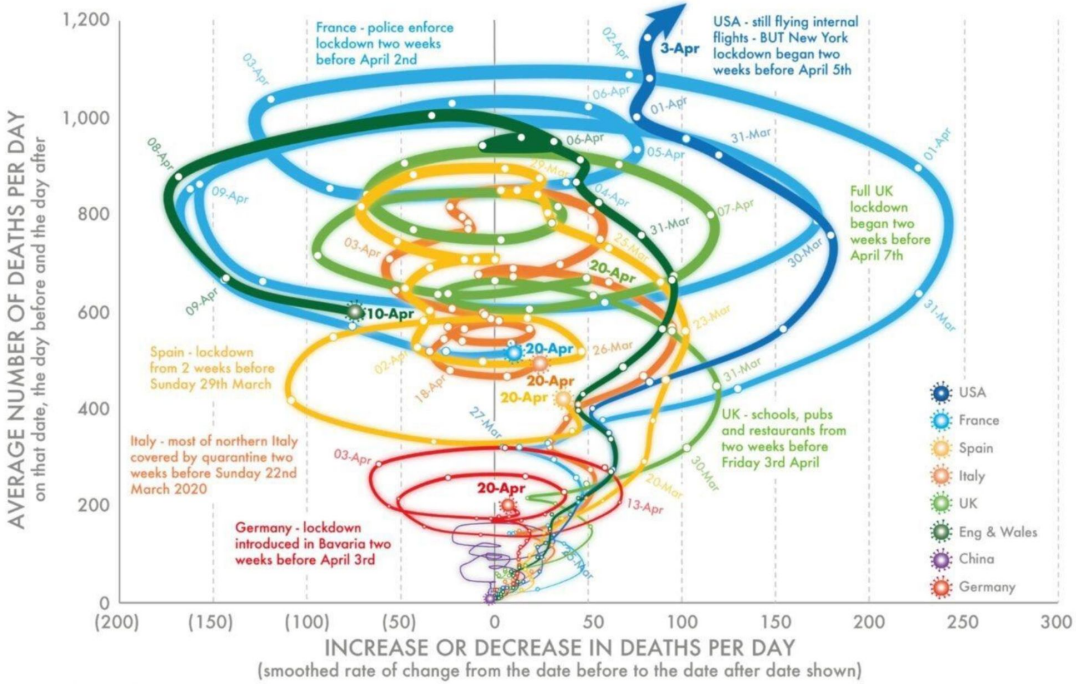


Fuente: Grupo Visioon. Test de daltonismo
https://visioon.es/blog/test-daltonismo-oftalmologia/

Fuente: Ayed, C. B., Halili, S., Tan, Y., & Grubb, A. M. (2023). Toward Internationalization and Accessibility of Color-based Goal Model Interpretation.

Visualización

Problemas **sobrecarga de información**



Visualización

Problemas generales

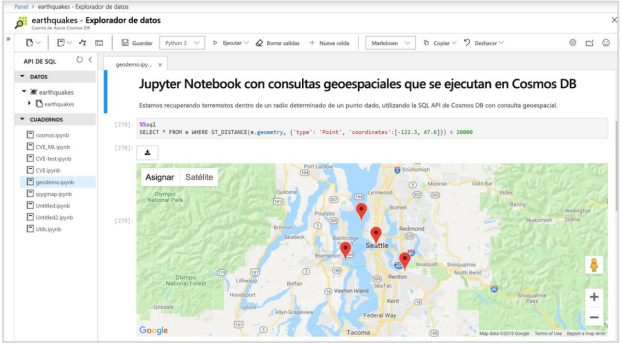
Via Reddit.
https://www.reddit.com/r/chile/comments/fpe6j5/megavisi%C3%B3n_being_very_megavisi%C3%B3n/

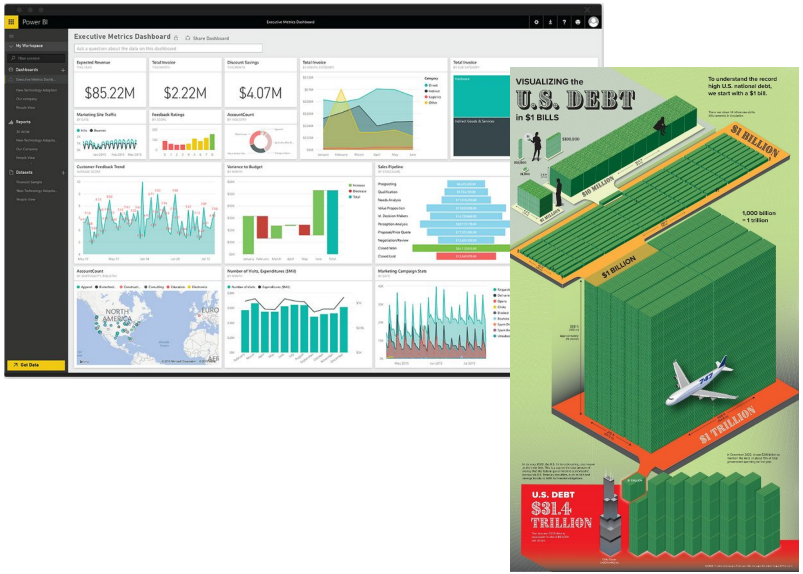
Visualización

Habitualmente es necesario utilizar múltiples visualizaciones para transmitir la información.

Podemos diferenciar las agrupaciones en:

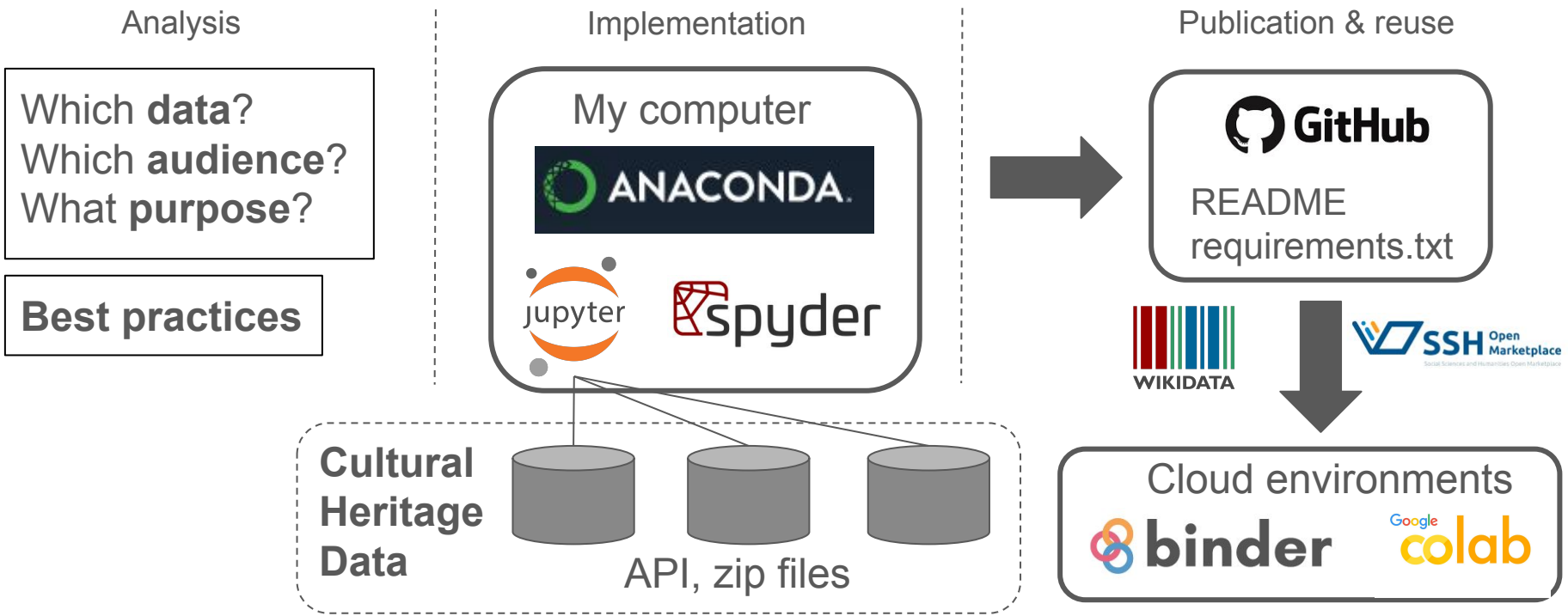
- **Notebooks**: incluyen narrativa y código
- **Cuadros de mando**: organizan las visualizaciones de forma analítica
- **Infografías**: comunican para concienciar sobre un tema particular





Visión por computador

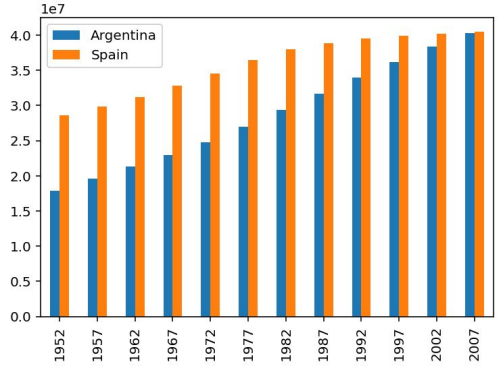
Herramientas - Jupyter Notebooks

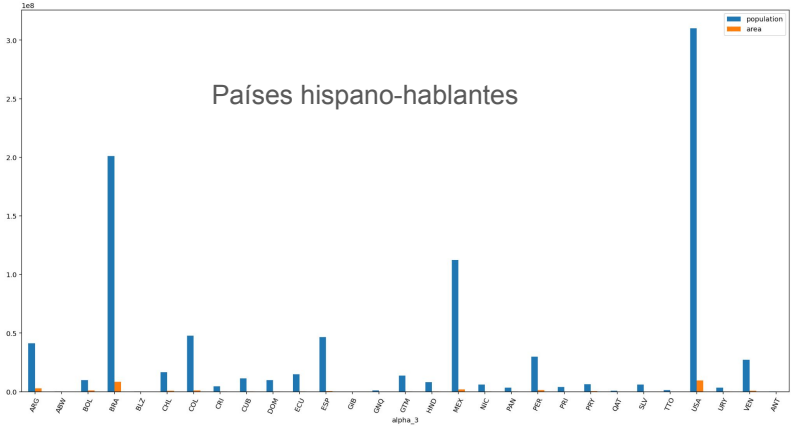


Visualización

Herramientas - Pandas

- Permite el análisis de datos tabulados (CSV, TSV)
- Importar, limpiar y gestionar datos
- Fusionar datos de forma eficiente
- Crear visualizaciones a partir de los datos





https://github.com/hibernator11/fi-ua

Visualización

Folium Herramientas

Python data, leaflet.js maps

Folium builds on the data wrangling strengths of the Python ecosystem and the mapping strengths of the Leaflet.js library. Manipulate your data in Python, then visualize it in a Leaflet map via Folium.

Concepts

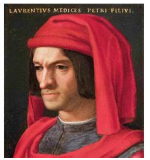
Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map.

The library has a number of built-in tilesets from OpenStreetMap, Mapbox, etc, and supports custom tilesets. Folium supports both Image, Video, GeoJSON and TopoJSON overlays and has a number of vector layers built-in.

Visualización

Herramientas - Wikidata

- Permite la creación de visualizaciones a partir de los datos de Wikidata
- Mapas, líneas de tiempo, cuadros de imágenes, etc.
- No es necesario instalar ningún software



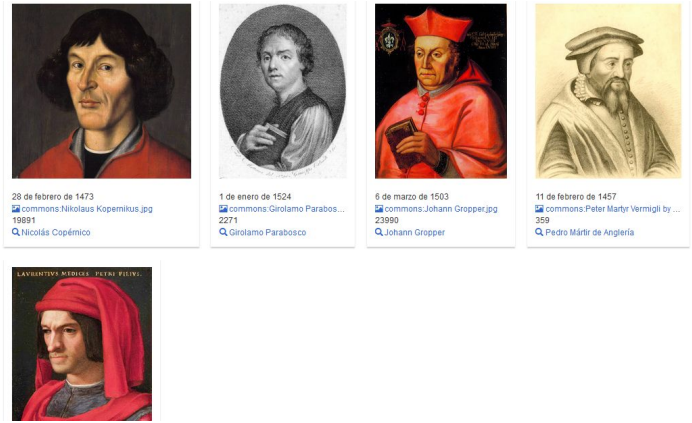
Visualización

Herramientas - Wikidata

- #defaultView:ImageGrid

https://w.wiki/8922

```
#defaultView:ImageGrid
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento .
  ?autor wdt:P18 ?imagen

  FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
         && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .

  SERVICE wikibase:label {
          bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es". }
}
LIMIT 10
```

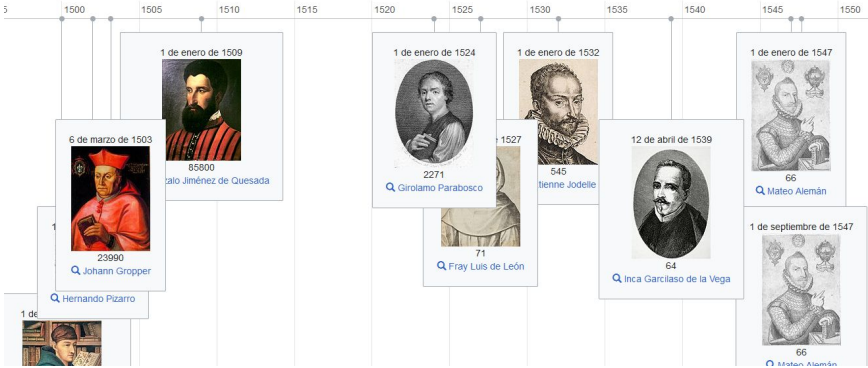


Visualización

Herramientas - Wikidata

- #defaultView:Timeline

https://w.wiki/88$y

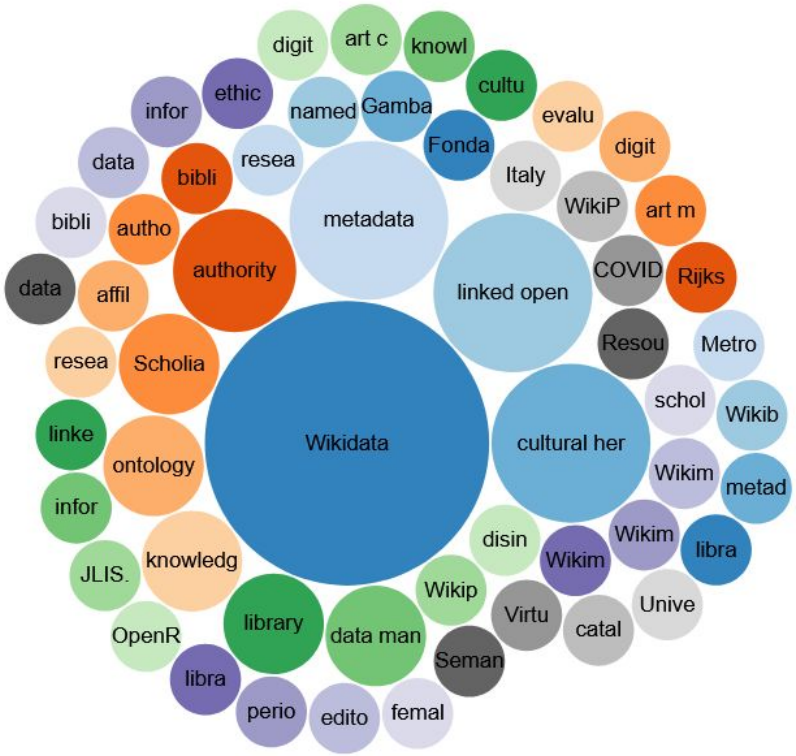



```
#defaultView:Timeline
SELECT ?imagen ?autor ?autorLabel ?idbvmc ?fechaNacimiento
WHERE {
  ?autor wdt:P2799 ?idbvmc .
  ?autor wdt:P569 ?fechaNacimiento .
  ?autor wdt:P18 ?imagen

  FILTER("1500-01-01"^^xsd:dateTime <= ?fechaNacimiento
         && ?fechaNacimiento < "1550-01-01"^^xsd:dateTime) .

  SERVICE wikibase:label {
          bd:serviceParam wikibase:language "[AUTO_LANGUAGE],es". }
}
LIMIT 10
```

MUSEO DEL PRADO | LÍNEA DEL TIEMPO
Busca un artista, una obra, un evento...
Contexto histórico
Federico I Barbarroja
Enrique II de Inglaterra
1133 - 1189
Jaime I de Aragón
Enrique III de Inglaterra
1207 - 1
Eduardo I de Inglaterra
1239 - 1307
Geoffrey Chaucer
Eduardo III de Inglaterra
1312
Guerra de los Cien Años
1337 - 1453
Francisco I de Francia
1494 - 1
Felipe II de España
Enrique VIII de Inglaterra
Carlos I de España
1500 - 1558
Carlos I de Inglaterra
1600 - 1649
Revolución inglesa
1643 - 1651
Thomas Jefferson
George Washington
1732 - 179
Abraham Lincoln
Napoleón Bonaparte
1769 - 1821
Primera Guerra
Guerra de Secesión
1861 - 1865
Segunda Guerra...
1939 - 1945
Frontal con
Frontal de
Cacería de
Parte superior del...
Anónimo
Frontal de
Retablo de San Cristóbal
Anónimo
Fantasía moral
La Creación
El Descendimie
Weyden, Rogier
Tríptico del Jardín de las...
El Bosco
Carlos V en la batalla de...
Tiziano Vecellio di
El caballero de la mano en el...
El Greco
Las tres
Las meninas
Velázquez, Diego R...
Bayeu y Subías,...
1734 - 1795
La maja
Goya y Lucientes,...
1746 - 1828
El 3 de mayo en Madrid o...
Goya y Lucientes, F...
Juana I
MUSEO DEL PRADO
Chicos en la playa
Sorolla y B
Piotr
Estud
cabez
Álvarez
https://www.museodelprado.es/coleccion/linea-del-tiempo
1100
1200
1300
1400
1500
1600
1700
1800
1900

Visualización

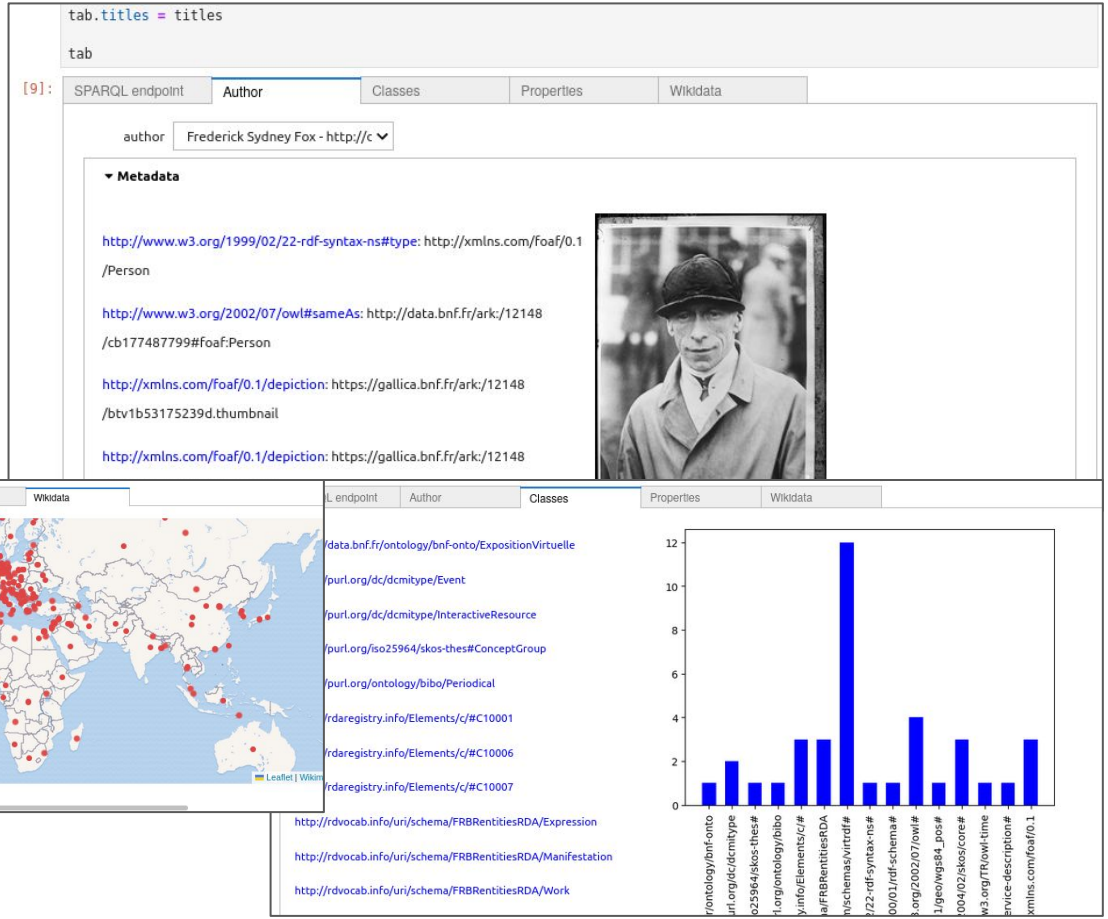
Herramientas - Wikidata



https://github.com/hibernator11/wikidata-review

Visualización

A **reproducible approach** to **browse** Linked Open Data in Cultural Heritage

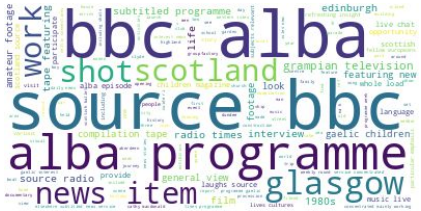


Schema.org

https://github.com/hibernator11/visual-ch-lod

https://doi.org/10.1145/3707647

Visualización



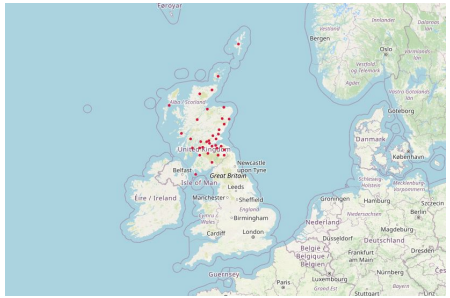
Moving Image Archive

This dataset represents the descriptive metadata from the Moving Image Archive catalogue, which is Scotland's national collection of moving images.

- Data format: metadata available as MARCXML and Dublin Core
- Data source: https://data.nls.uk/data/metadata-collections/moving-image-archive/

The Jupyter Notebooks include a set of examples to reproduce the transformation to RDF and enrichment with external repositories:

- Data extraction
- Exploring the CSV text file
- Transformation to LOD
- Enrichment
- Exploring with SPARQL
- Exploring geographic locations
- Data Quality assessment



jupyter

Folium

WIKIDATA

MARCXML MARC 21 XML Schema Official Web Site

Schema.org

GeoNames

WIKIDATA

Visualización

Proyecto presentado en la actividad "Compartir pantalla" con la Asociación de Humanidades Digitales Hispánicas.

launch binder

DOI 10.5281/zenodo.10123566

HDH Humanidades Digitales Hispánicas

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

Universitat d'Alacant
Universidad de Alicante

Introducción

Este proyecto forma parte de la actividad "Compartir pantalla" que introduce Wikidata en el ámbito bibliotecario como plataforma para facilitar el acceso, enriquecimiento y la visibilidad a través de la edición de recursos y el uso de SPARQL para recuperar información en forma de datos y recursos gráficos.

Además, se incluyen ejemplos de sentencias SPARQL proporcionados por Wikidata introduciendo los diferentes comandos e instrucciones, visualizaciones reales utilizadas en la web de Biblioteca Virtual Miguel de Cervantes y la International GLAM Labs Community, y otros ejemplos ilustrativos como por ejemplo sentencias federadas y creación de datasets.

https://github.com/hibernator11/hdh-compartir-pantalla-2023

Infraestructuras para el almacenamiento y procesamiento de datos

Adquisición y preparación de datos

IA EPS UA

Universitat d'Alacant
Universidad de Alicante

Contenido

- Introducción
  - Ciencia Abierta y Digital Commons
  - Enlaces permanentes
  - Colecciones como datos
  - Principios FAIR y CARE
  - Documentación (datasheets)
- Infraestructuras
  - Zenodo y GitHub
  - Wikibase Cloud
  - Espacios de datos
  - EOSC y ECCCH

Economía del dato

En **2025**, se espera que las iniciativas europeas sobre datos y regulación de datos evidencien resultados: se espera un fuerte aumento del valor de la economía del dato como parte de la economía general, con un aumento del impacto de la economía del dato en términos de PIB de 1 punto porcentual.

Asimismo, en **2030**, se espera que la economía del dato (UE27 + Reino Unido) supere el 1 billón de euros, con una tasa de crecimiento anual del 5,6 % entre 2025 y 2030.

MINISTERIO DE ECONOMÍA, COMERCIO Y EMPRESA

Oficina del **Dato**

https://datos.gob.es/sites/default/files/doc/file/OdD/Economiadeldato.pdf

Economía del dato

Otra contribución importante a este crecimiento de la economía del dato son las medidas de inversión integradas en los **fondos Next Generation EU** y las reformas nacionales de recuperación y resiliencia, que permitirán que los países implementen una serie de inversiones en torno a tecnologías clave, desde tecnología cloud hasta 5G, IoT y dispositivos inteligentes, inteligencia artificial, e infraestructura digital, entre otras.

MINISTERIO DE ECONOMÍA, COMERCIO Y EMPRESA

Oficina del Dato

https://datos.gob.es/sites/default/files/doc/file/OdD/Economiadeldato.pdf

Datos en Europa

Los datos es un recurso esencial para el crecimiento económico, competitividad, innovación, la creación de empleo y el progreso de la sociedad en general. En el futuro, el desarrollo de aplicaciones de datos traerá beneficios para los ciudadanos y las empresas:

- mejorar la salud
- crear sistemas de transporte más seguros y limpios
- generar nuevos productos y servicios
- reducir el coste de los servicios públicos
- mejorar la sostenibilidad y la eficiencia energética

https://digital-strategy.ec.europa.eu/en/policies/strategy-data

Infraestructuras

La mayoría de datos disponibles son almacenados por grandes corporaciones y centros de datos en la nube como Amazon Web Services (AWS), Google Cloud Platform (GCP), y Microsoft Azure.

Actualmente existen problemas a la hora de la transparencia y reuso de datos con fines económicos como el entrenamiento de modelos de IA privados. Numerosas instituciones han promovido y restringido el acceso a sus colecciones:

- We restrict access of commercial AI to KB collections

En este nuevo contexto, la Unión Europea ha apostado por un modelo federado de compartición de datos denominado espacio de datos. Un espacio de datos (en inglés, *dataspace*) es el lugar de generación de valor alrededor del dato mediante su compartición voluntaria en un entorno de soberanía, confianza y seguridad.

https://datos.gob.es/es/blog/por-que-espacios-de-datos

datos.gob.es
reutiliza la información pública

Ciencia Abierta - Digital Commons

La **Ciencia Abierta (Open Science)** es un movimiento internacional para facilitar el acceso a los datos abiertos, software con licencias open source, métodos, resultados y la colaboración.

- https://creativecommons.org/about/open-science/
- https://research-and-innovation.ec.europa.eu/strategy/strategy-research-and-innovation/our-digital-future/open-science_en

Digital Commons o “Bienes comunes digitales” se refiere a los **recursos digitales compartidos, como software, conocimiento y datos**, que son producidos y gestionados colectivamente por una comunidad, con el objetivo de que sean de uso público.

Se caracterizan por el **acceso abierto, la gestión participativa y las prácticas de licencia que preservan la reutilización y redistribución**. Provee la infraestructura y plataformas para implementar las prácticas de Ciencia Abierta.

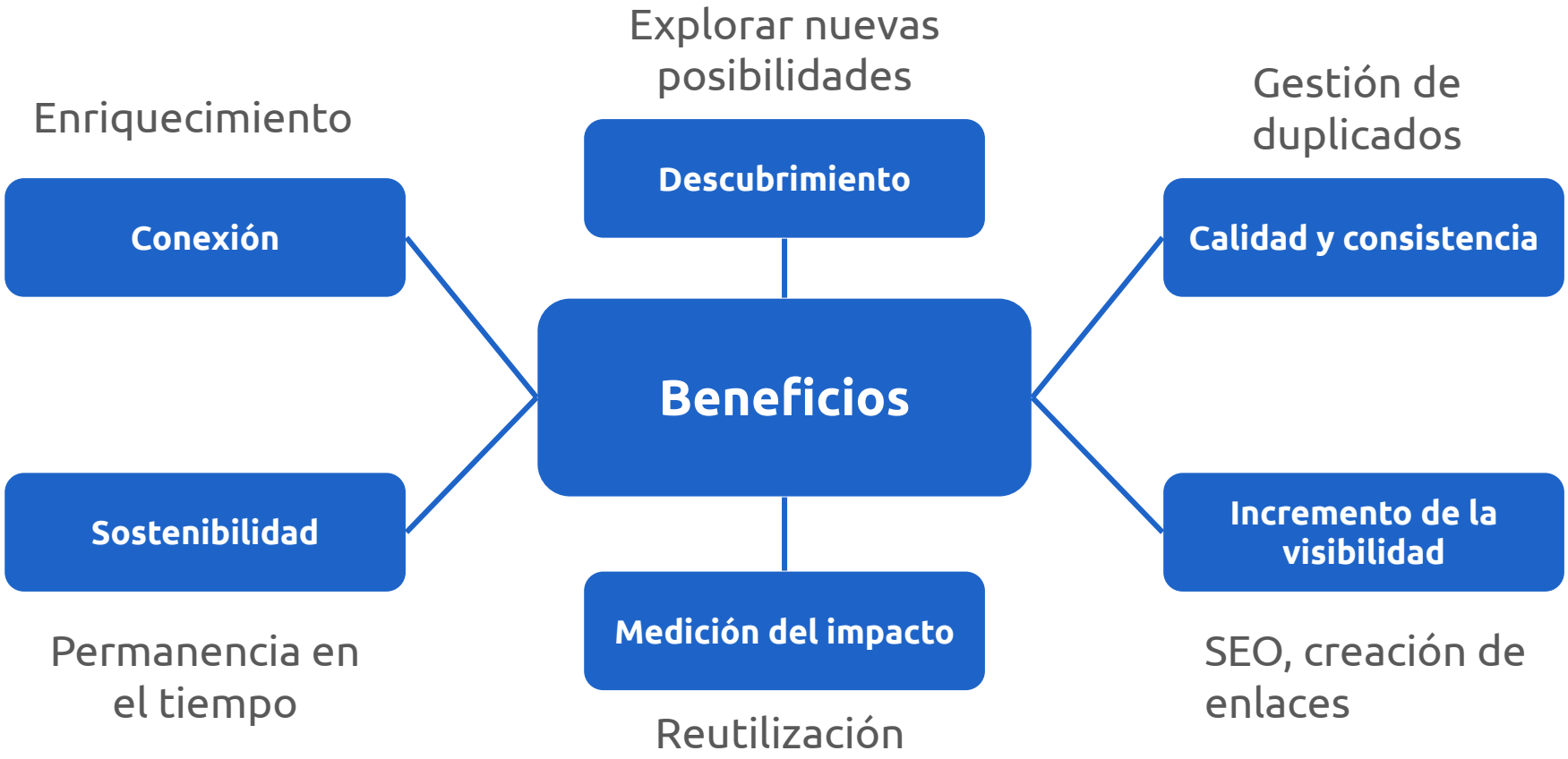
¿Qué es un enlace permanente?

- Permite a un usuario encontrar un contenido (documento, web o recurso) a través de una **URL que siempre estará accesible**
- Algunos **ejemplos** son Orcid o DOI
- En ocasiones, facilitan información sobre el **tipo de recurso**
  - https://data.cervantesvirtual.com/person/40
- Mejora la gestión de la información (duplicados?)
- Facilita la adopción de tecnologías avanzadas como datos enlazados y la Web Semántica

¿Qué es un enlace permanente?

- Numerosas **iniciativas** están relacionadas con la definición de buenas prácticas y estrategias para la **creación y gestión de enlaces permanentes**
  - FAIR principles
  - The British Library Adopts a New Persistent Identifier Policy
  - Europeana - Policy for persistent identifiers in the data space
  - National Information Standards Organization - National PID Strategies
  - Research Data Alliance - National PID Strategies WG
  - Politique identifiants BnF
  - Australian National PID Strategy and Roadmap
  - Code4lib - Persistent identifiers for heritage objects

Enlaces permanentes



WIKIPEDIA
La enciclopedia libre

juel de Cervantes en la Biblioteca Digital Hispánica
de Cervantes en la Biblioteca Digital Memoriademadrid.

Control de autoridades	**Proyectos Wikimedia** · Datos: Q5682 · Multimedia: Miguel de Cervantes / Q5682 · Citas célebres: Miguel de Cervantes · Textos: Autor:Miguel de Cervantes **Identificadores** · WorldCat · VIAF: 17220427 · ISNI: 0000000121221919 · BNA: 000031582 · BNA: 000031582 · BNE: XX1718747 · BNF: 118957747 (data) · VcBA: 495/19174 · BNC: 000003096 · BNC: 000003096 · BNB: 000155490 · CANTIC: 981058519886006706 · GND: 11851993X · LCCN: n79100233 · NDL: 00435606 · NKC: jn19990001383 · NLA: 36587208 · BER: 000001810 · NLI: 987007259788205171 · CiNii: DA00334243 · SNAC: w6g15z8k · SUDOC: 026774771 · ULAN: 500371871 · BIBSYS: 90055928 · ICCU: CFIV008229 · ARAE: cervantes-saavedra-miguel-de-1547-1616 · Deutsche Biographie: 11851993X · Open Library: OL676237A · KulturNav: id · **Diccionarios y enciclopedias** · DBE: url · Britannica: url · ELEM: 3332 · Treccani: url · **Repositorios digitales** · BVMC: 40 · Dialnet: 2995919 · Proyecto Gutenberg: 505 · BDCYL: 45511 · BVPB: 23844 · BiValDi: 337 · Europeana: agent/base/60157 · **Patrimonio histórico** · FMIS: LSH/agents/66829 · **Cine** · IMDb: nm0148859

Categorías: Hombres | Nacidos en 1547 | Fallecidos en 1616 | Miguel de Cervantes | Escritores de Alcalá de Henares | Castellanos | Escritores de España del Siglo de Oro

Apariel
Texto
Pe
Es
Gr
Anchur
Es
An
Color (
Au
Cl
Os

https://es.wikipedia.org/wiki/Miguel_de_Cervantes

Ejemplos de enlaces permanentes

- https://data.cervantesvirtual.com/person/40
- https://data.cervantesvirtual.com/work/766798
- https://datos.bne.es/persona/XX1718747.html
- https://datos.bne.es/obra/XX2348547.html
- https://data.bnf.fr/11928669/voltaire/
- https://id.loc.gov/resources/works/10118761.html
- http://www.wikidata.org/entity/Q5682

El tipo del recurso se incluye en la URL

En otros casos no se incluye el tipo

Dominio + Tipo de entidad + Identificador

Ejemplos de enlaces permanentes

BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

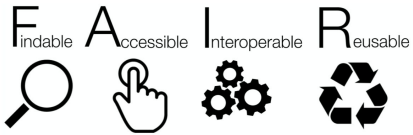
https://data.cervantesvirtual.com

- https://data.cervantesvirtual.com/person/40
- https://data.cervantesvirtual.com/corporatebody/7802
- https://data.cervantesvirtual.com/work/766798
- https://data.cervantesvirtual.com/expression/766799
- https://data.cervantesvirtual.com/manifestation/766800
- https://data.cervantesvirtual.com/item/134373

Dominio + Tipo de entidad + Identificador

Collections as Data

https://collectionsasdata.github.io



Características de los datos

Findable
Rich metadata and permanent identifiers

Accessible
machine-readable metadata

Interoperable
Metadata described using vocabularies

Reusable
Licence and documentation

Complementa FAIR considerando el beneficio colectivo y el uso responsable de los datos

Collective benefit
Inclusive and equitable outcomes

Authority to control
Recognizing rights and interests

Responsibility
For indigenous languages and worldviews

Ethics
Minimizing harm, and for justice and future use

Collections as Data

europeana pro

https://pro.europeana.eu/index.php/page/harvesting-and-downloads

Repositorios de datos abiertos y enlazados

SAAM

https://americanart.si.edu/about/lod

Data Foundry

Data collections from the National Library of Scotland

https://doi.org/10.5281/zenodo.8051036

LIBRARY LABS

LIBRARY OF CONGRESS

https://id.loc.gov

KUNGL. BIBLIOTEKET

National Library of Sweden

https://libris.kb.se/sparql

1640 THE NATIONAL LIBRARY OF FINLAND

https://data.nationallibrary.fi

BRITISH LIBRARY

https://bnb.data.bl.uk

DATOS·BNE·es

https://datos.bne.es/

BnF Data

https://data.bnf.fr/

KB LAB

https://data.bibliotheken.nl

CENTRO BIBLIOTECA VIRTUAL MIGUEL DE CERVANTES

https://data.cervantesvirtual.com/

Bibliothèque nationale du Luxembourg

Open Data

https://data.bnl.lu/

DEUTSCHE NATIONAL BIBLIOTHEK

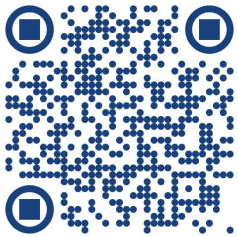
https://www.dnb.de/EN/lds

Österreichische Nationalbibliothek

https://labs.onb.ac.at/en/dataset/lod

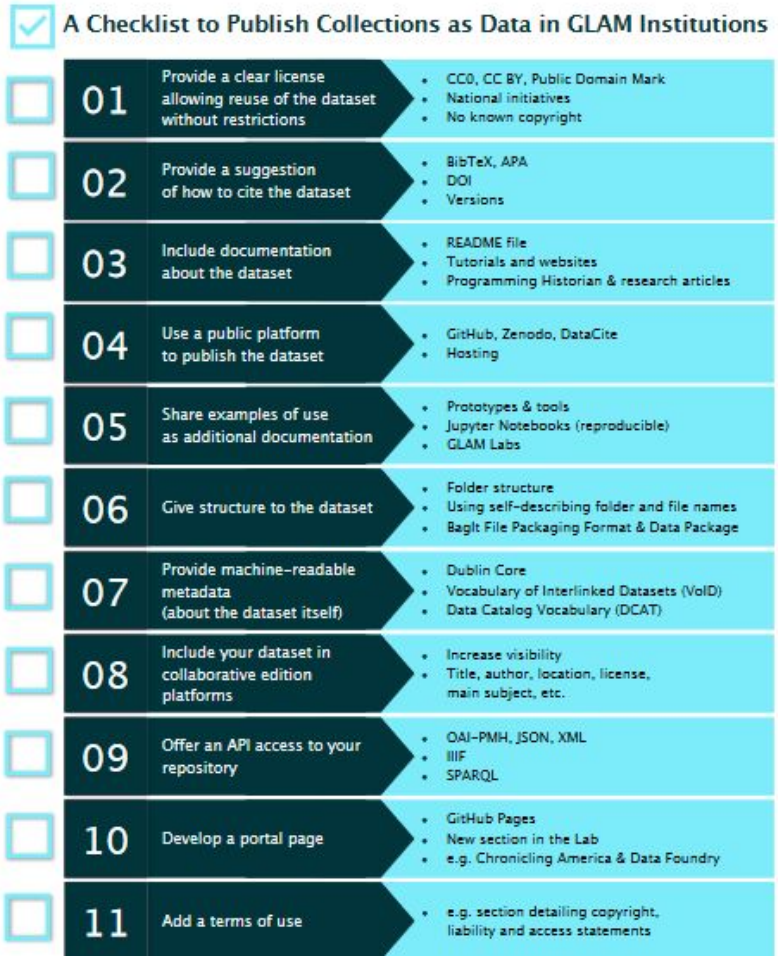
Collections as Data

A checklist to make available digital collections suitable for computational use



https://doi.org/10.48550/arXiv.2304.02603

Candela, G., Gabriëls, N., Chambers, S., Dobreva, M., Ames, S., Ferriter, M., Fitzgerald, N., Harbo, V., Hofmann, K., Holownia, O., Irollo, A., Mahey, M., Manchester, E., Pham, T.-A., Potter, A. and Van Keer, E. (2025), "A checklist to publish collections as data in GLAM institutions", Global Knowledge, Memory and Communication, Vol. 74 No. 5/6, pp. 1323-1355. https://doi.org/10.1108/GKMC-06-2023-0195



https://glamlabs.io/checklist/

SSHOpenCluster
Social Sciences & Humanities Resources

Collections as Data Workflow

SSH Open Marketplace

Report an issue | Sign in

Tools & services | Training materials | Publications | Datasets | Workflows | Browse | Contribute | About

Search

Home / Workflows / A workflow to publish Collections as Data: the case of Cultural Heritage data spaces

A workflow to publish Collections as Data: the case of Cultural Heritage data spaces

Copy to clipboard

Cultural Heritage institutions have been making their digital collections available for the public for several decades. Advances in technology such as Artificial Intelligence and Machine Learning have provided a new context in which digital collections can be analysed using computational methods. Initiatives such as Collections as data and the FAIR data principles, have emerged to provide best practices and guidelines for publishing digital collections suitable for computational use. These initiatives are complemented with the CARE principles to strengthen ethical considerations in data governance and reuse. In parallel, experimental Labs have been implemented in Galleries, Libraries, Archives and Museums (GLAM) in order to reuse the digital collections.

Data spaces have emerged as an innovative way to publish and reuse digital collections. Based on previous work undertaken in the context of the GLAM Labs Community, this workflow provides a set of steps to publish Collections as data. It aims to guide and encourage cultural heritage institutions, step by step, in publishing their collections, so that they are suitable for computational use. It is important

Details

ACCESS

License Creative Commons Attribution 4.0 International

CATEGORISATION

Activity Description Extracting Analyzing Discovering Gathering Disseminating Sharing

Keyword digital collections Data computational methods Collections as data

Candela, G., Chambers, S. and Irollo, A. (2023) A workflow to publish Collections as Data: the case of Cultural Heritage data spaces.
https://marketplace.sshopencloud.eu/workflow/I3JvP6

COMMON EUROPEAN DATA SPACE FOR CULTURAL HERITAGE

europeana

COLLECTIONS AS DATA

DARIAH-EU
Digital Research Infrastructure for the Arts and Humanities

Workflow steps (10)

1. Provide a clear license and terms of use allowing reuse of the dataset without restrictions — Expand
2. Provide a suggested citation for the dataset so reusers are aware of how to cite it — Expand
3. Include documentation about the dataset — Expand
4. Use a public platform to make available the dataset for the public — Expand
5. Share examples of use to demonstrate how the dataset can be reused
6. Think about a structure for the dataset for a better understanding of how to reuse the content — Expand
7. Include machine-readable metadata about the content provided in the dataset — Expand
8. Use an existing collaborative-edition platform to include the information about the dataset — Expand
9. Provide the dataset by means of an existing API — Expand
10. Create a website to present and describe the dataset to encourage its reuse — Expand

Documentación (Datasheets)

Datasheets for cultural heritage datasets

This Working Group, set up within the Europeana Research Community and EuropeanaTech Community, works to adapt the concept of datasets for the cultural heritage sector.

Journal of Open Humanities Data

Reading: Datasheets for Digital Cultural Heritage Datasets

Discussion Paper

Datasheets for Digital Cultural Heritage Datasets

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudecker, Giulia Osti, Daniel van Strien

europeana

https://doi.org/10.5334/johd.124

https://pro.europeana.eu/project/datasheets-for-digital-cultural-heritage-working-group

Documentación (Datasheets)

https://zenodo.org/records/13284442

zenodo | Search records... | Communities | My dashboard | gcandela...

Datasets of Staatsbibliothek zu Berlin - Berlin State Library

Published August 9, 2024 | Version v2

Dataset | Open

Metadata of the "Alter Realkatalog" (ARK) of Berlin State Library (SBB)

Lehmann, Jörg (Data curator)[1]; Schneider, Sophie (Data curator)[1]

Show affiliations

This dataset was created with the intent to provide a single larger set of metadata from Berlin State Library for research purposes and the development of AI applications.

The dataset comprises of descriptive metadata of 2.619.397 titles, which together form the "Alte Realkatalog" of Berlin State Library, which may be translated to "Old Subject Catalogue". The data are stored in columnar format, containing 375 columns. They were downloaded in December 2023 from the German central library system (CBS). Exemplary tasks which can be served by this dataset comprise studies on the history of books between 1500 and 1955, on the paratextual formatting of scientific books between 1800 and 1955, and on pattern recognition on the basis of bibliographical metadata.

Files

Metadata of the "Alter Realkatalog" (ARK) of Berlin State Library (SBB)-V2.md

Metadata of the "Alter Realkatalog" (ARK) of Berlin State Library (SBB)

Motivation

This dataset was created with the intent to provide a single larger set of metadata from Berlin State Library for research purposes and the development of AI applications.

The dataset comprises of descriptive metadata of 2.619.397 titles, which together form the "Alte Realkatalog" of Berlin State Libray, which may be translated to "Old Subject Catalogue". The data are stored in columnar format, containing 375 columns. They were downloaded in December 2023 from the German central library system (CBS). Exemplary tasks which can be served by this dataset comprise studies on the history of books between 1501 and 1955, on the paratextual formatting of scientific books between 1800 and 1955, and on pattern recognition on the basis of bibliographical metadata.

3K VIEWS | 371 DOWNLOADS

Show more details

Versions

Version v2 — Aug 9, 2024
10.5281/zenodo.13284442

Version 1 — Jul 22, 2024
10.5281/zenodo.12783814

View all 2 versions

Cite all versions? You can cite all versions by using the DOI 10.5281/zenodo.12783813. This DOI represents all versions, and will always resolve to the latest one. Read more.

External resources

Indexed in

OpenAIRE

Communities

Datasets of Staatsbibliothek zu Berlin - Berlin

Github

hibernator11

Type / to search

Overview | Repositories 66 | Projects | Packages | Stars

Find a repository… | Type | Language | Sort | New

Gustavo Candela
hibernator11

Gustavo Candela is a Lecturer in Computer Science at the University of Alicante.

Edit profile

24 followers · **14** following

University of Alicante
Alicante
@gus_candela

Organizations

wikidata-queries Public
examples of Wikidata SPARQL queries
Jupyter Notebook Updated 3 days ago
Star

Spanish-Civil-War-KGs Public
Extracting Knowledge Graphs as Collections as Data using the Spanish Civil War as main topic
Jupyter Notebook Updated last week
Star

KU-Leuven-Libraries-Impact-CollectionsAsData Public
KU-Leuven-Libraries-Impact-CollectionsAsData
Jupyter Notebook Updated 3 weeks ago
Star

dhwiki-notebooks Public
dhwiki Jupyter notebook examples
Jupyter Notebook Updated on Jul 8
Star

workshop-notebooks-wac-dhnb2025 Public
Examples of code to extract Web Archive Content to create collections as data
Jupyter Notebook 1 Updated on May 4
Star

Zenodo

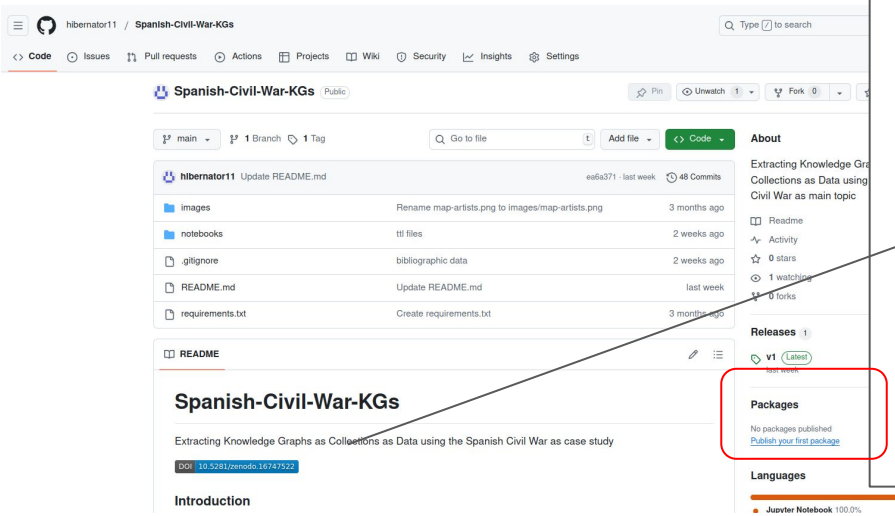
Zenodo es un repositorio de acceso abierto de propósito general desarrollado bajo el programa europeo OpenAIRE y operado por CERN.

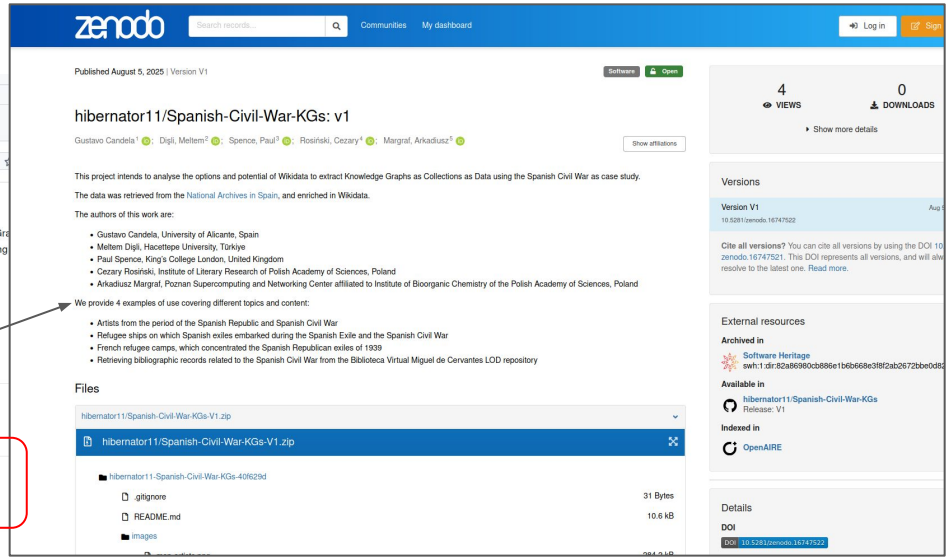
Permite a investigadores de todas las disciplinas depositar y compartir resultados de investigación, como artículos, datos, software, informes y otros tipos de artefactos digitales, en diversos formatos.

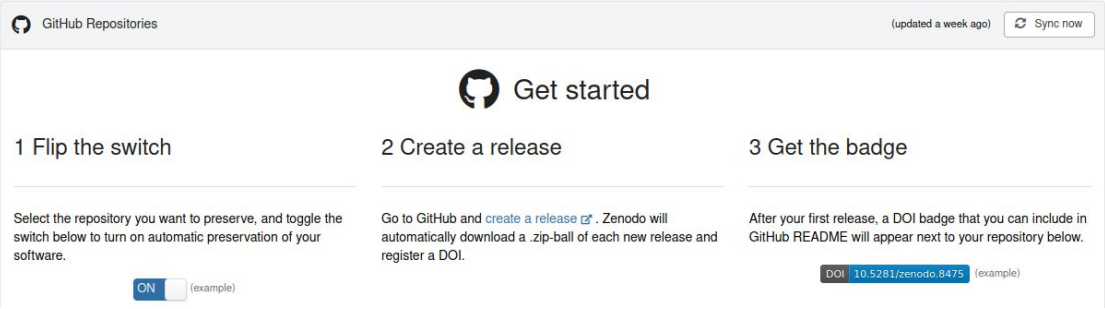
Cada envío recibe un DOI (Digital Object Identifier), lo que facilita su localización y citación.

https://zenodo.org/

Zenodo

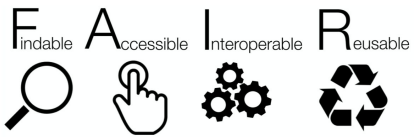






Integrar Zenodo (doi) y Github (versión-release)

Wikibase



Software de **código abierto** que se usa en Wikidata.

Útil para crear bases de conocimientos a través de los **principios FAIR y Linked Open Data**.

Permite crear tu propio modelo de datos (propiedades). Dispone de API y agregadores de datos.

Servicio disponible en la nube. Permite crear hasta 6 instancias y configurar el acceso y diseño.

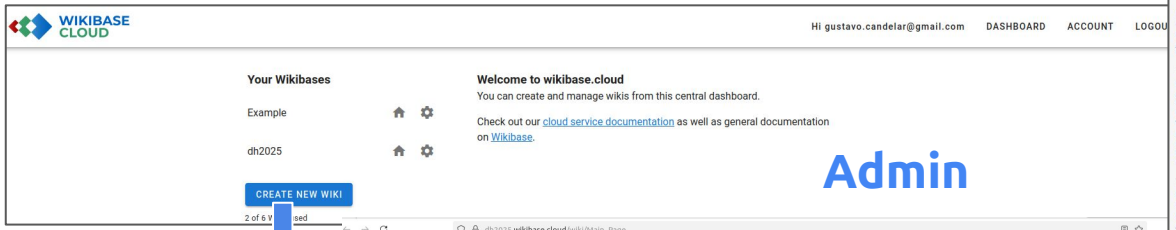
WIKIBASE

https://wikiba.se/

WIKIBASE CLOUD

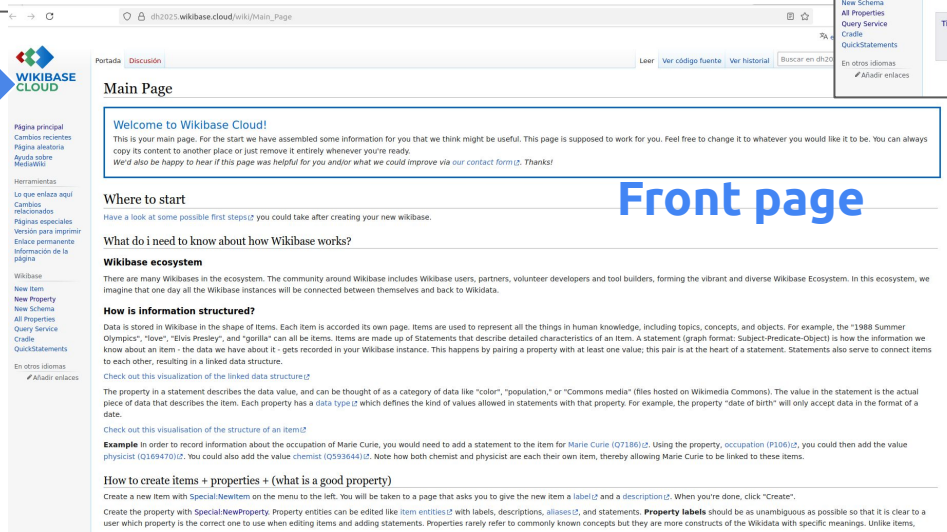
https://www.wikibase.cloud/

Wikibase

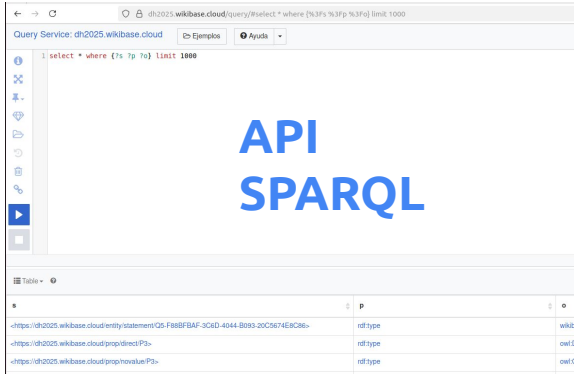


Admin

Entidades



Front page



API
SPARQL

https://dh2025.wikibase.cloud/wiki/Main_Page

DARIAH-EU WG DHwiki

Post on dariah.eu

A space for discussion and dissemination of use case experiences and best practices around **Wikibase and Wikidata in a DH context**, and for contributions to further developing the Wikibase software and related tools.

Building bridges between sectors

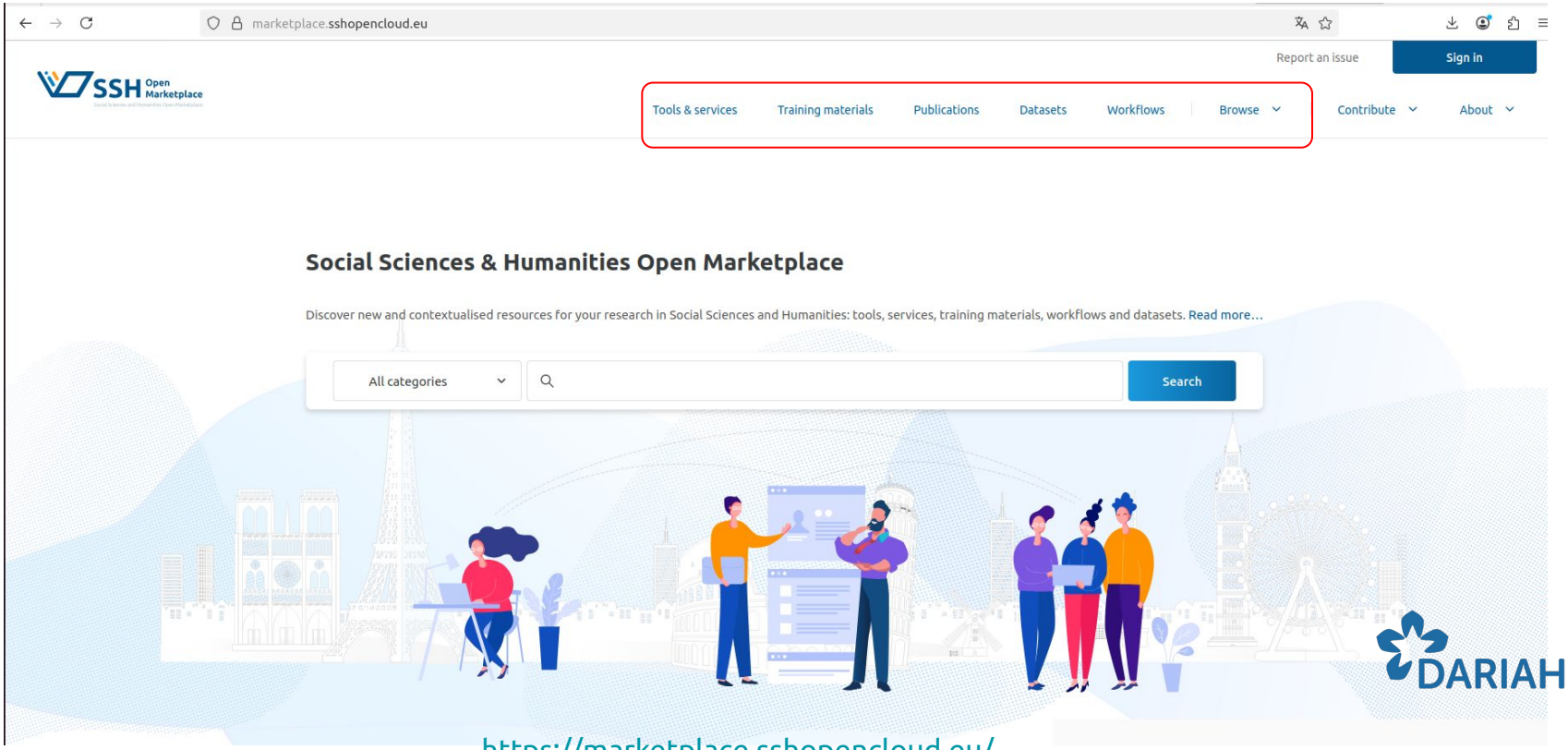
- GLAM
- DH research
- Wikimedia affiliates / platform users

Image source

DHWIKI
A DARIAH-EU WORKING GROUP

DARIAH-EU
Digital Research Infrastructure
for the Arts and Humanities

Social Sciences and Humanities Open marketplace



https://marketplace.sshopencloud.eu/

Infraestructuras

Espacio de datos es un entorno seguro de compartición de datos federado donde las empresas e instituciones mantienen la soberanía de los datos

Actores y participantes en espacios de datos

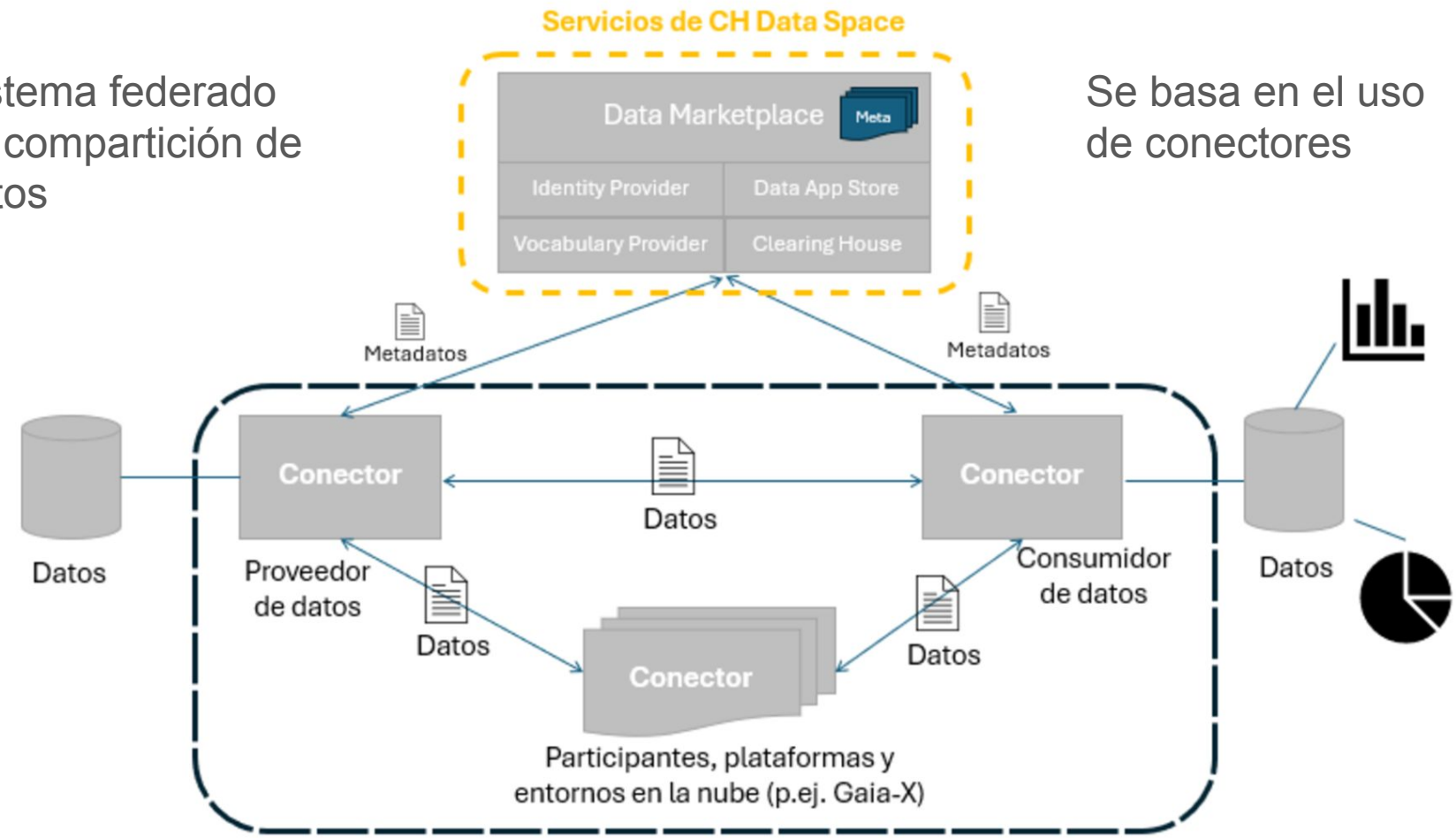
- promotor: responsable gobierno y gestión
- proveedores de conjuntos y servicios de datos
- consumidores de conjuntos y servicios de datos
- proveedor tecnológico: desplegar el espacio de datos para su uso

https://datos.gob.es/es/blog/por-que-espacios-de-datos

datos.gob.es
reutiliza la información pública

Sistema federado de compartición de datos

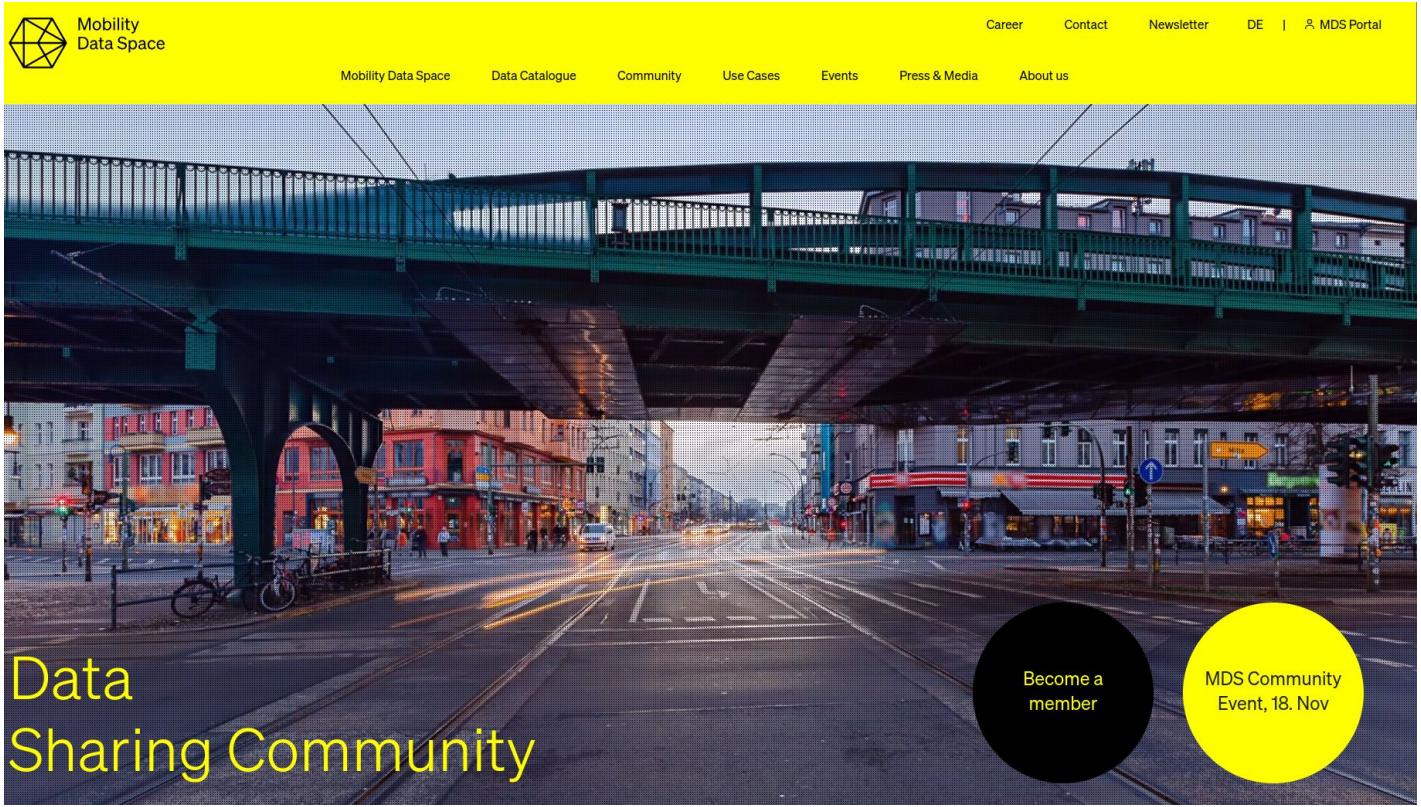
Se basa en el uso de conectores



Infraestructuras

https://www.dataspace-culturalheritage.eu/en

Infraestructuras



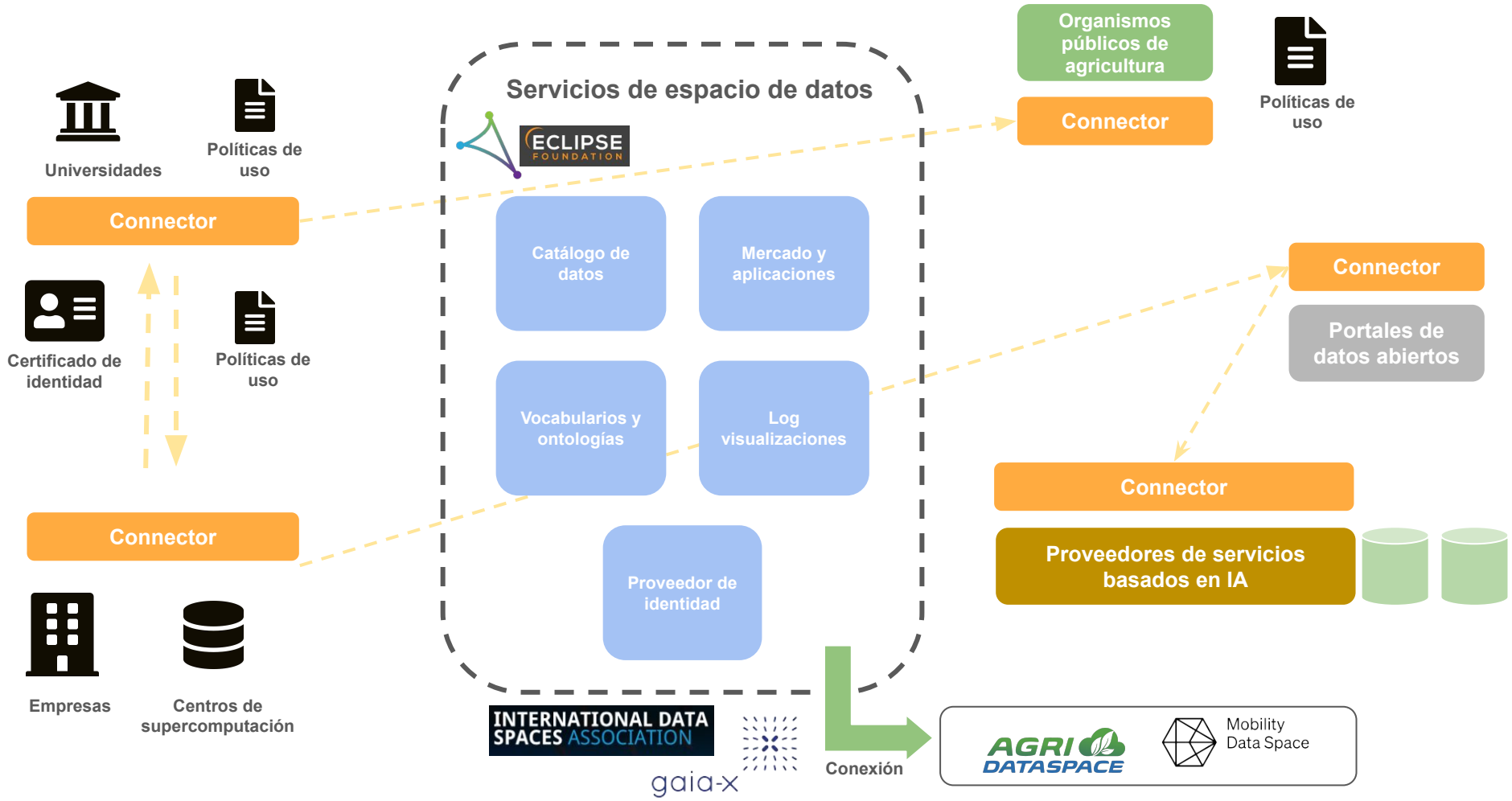
Inicio Kit espacio de Datos Proyecto Equipo Ética y seguridad Noticias Recursos Contacto

Español

OmicSpace impulsa la colaboración entre instituciones, conecta datos clínicos y ómicos de forma segura

https://omicspace.iislafe.es/

https://agrovalai.lucentia.es/


Universidades
Políticas de uso
Connector
Certificado de identidad
Políticas de uso
Connector
Empresas
Centros de supercomputación
Servicios de espacio de datos
ECLIPSE FOUNDATION
Catálogo de datos
Mercado y aplicaciones
Vocabularios y ontologías
Log visualizaciones
Proveedor de identidad
Organismos públicos de agricultura
Connector
Políticas de uso
Connector
Portales de datos abiertos
Connector
Proveedores de servicios basados en IA
INTERNATIONAL DATA SPACES ASSOCIATION
gaia-x
Conexión
AGRI DATASPACE
Mobility Data Space

AGROVAL-AI

Promotor Universidad de Alicante

INTERNATIONAL DATA SPACES ASSOCIATION

gaia-x

Proveedores de servicios

ÁGORA

UNIVERSIDAD DE MÁLAGA

Proveedores de datos

weitec

AEMet Agencia Estatal de Meteorología

datos.gob.es reutiliza la información pública

Cincofresh PISTACHOS

Ayuntamiento de Alicante

Vinícola de Tomelloso

AJUNTAMENT DE VALÈNCIA

UNIVERSIDAD DE CÓRDOBA

UCLM UNIVERSIDAD DE CASTILLA-LA MANCHA

Consumidores de datos

GENERALITAT VALENCIANA

Usuarios

Instituciones

DS y centros demostradores

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

UPCXels

AGRI DATASPACE

CeDEDA Centro Demostrador del Espacio de Datos Agroalimentario de Andalucía

Junta de Andalucía Consejería de Agricultura, Pesca, Agua y Desarrollo Rural

AgrospAI

Universitat de Lleida

Centro Demostrador de espacio de datos agroalimentarios

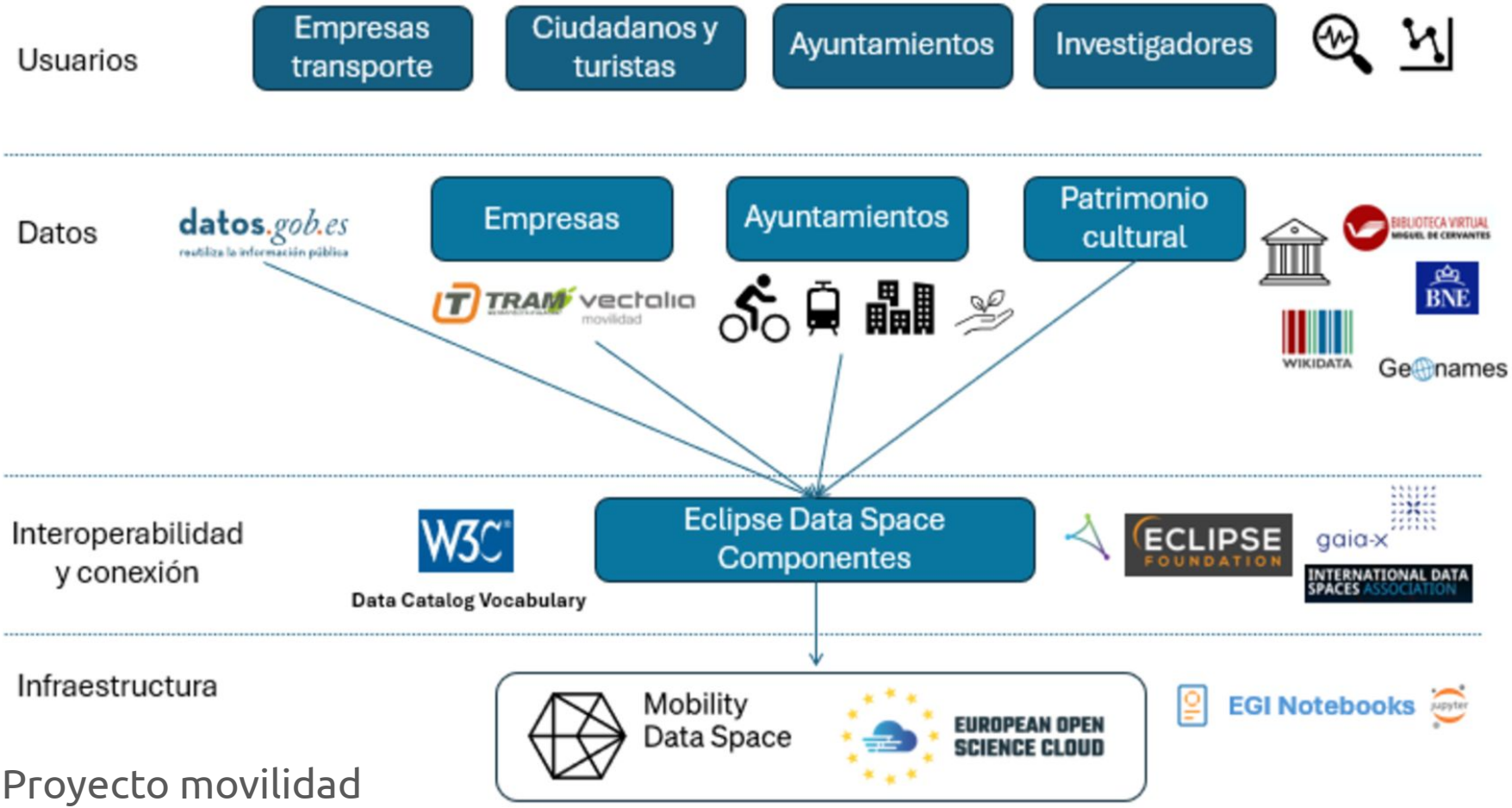
CTIC

seresco

Centro Demostrador de Espacio de datos Agroalimentario

Infraestructuras

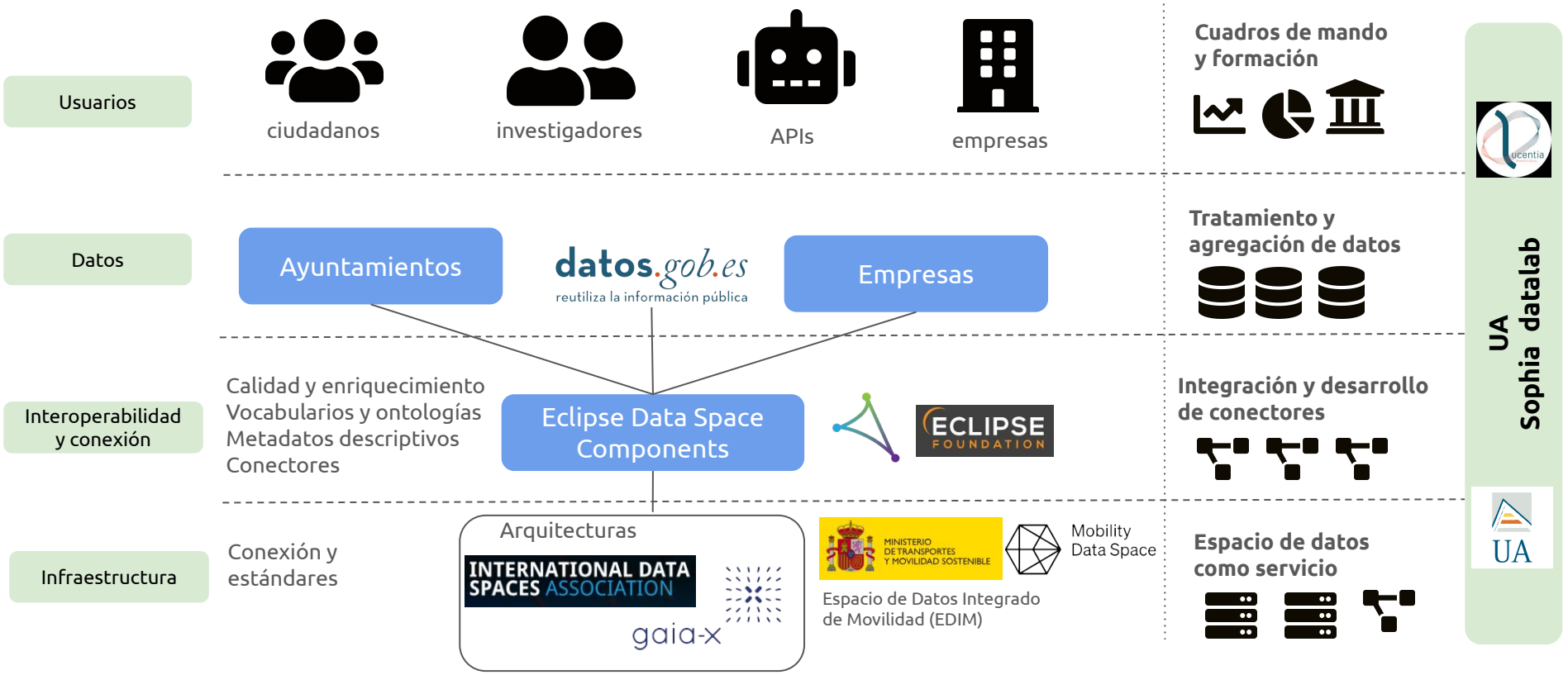
https://movilidad.lucentia.es/



Proyecto movilidad

Infraestructuras

https://sophia.lucentia.es/


Usuarios
ciudadanos
investigadores
APIs
empresas
Cuadros de mando y formación
Datos
Ayuntamientos
datos.gob.es
reutiliza la información pública
Empresas
Tratamiento y agregación de datos
Interoperabilidad y conexión
Calidad y enriquecimiento
Vocabularios y ontologías
Metadatos descriptivos
Conectores
Eclipse Data Space Components
ECLIPSE FOUNDATION
Integración y desarrollo de conectores
Infraestructura
Conexión y estándares
Arquitecturas
INTERNATIONAL DATA SPACES ASSOCIATION
gaia-x
MINISTERIO DE TRANSPORTES Y MOVILIDAD SOSTENIBLE
Mobility Data Space
Espacio de Datos Integrado de Movilidad (EDIM)
Espacio de datos como servicio
UA
Sophia datalab
UA

European Open Science Cloud (EOSC)

El nodo de la UE EOSC adquirido tiene las siguientes propuestas de valor:

- Facilita la creación de la «Web de datos FAIR y servicios interoperables» (denominada Federación EOSC) en el marco de la política de ciencia abierta.
- Define la ruta y el plan (es decir, el marco de arquitectura e interoperabilidad de EOSC) para que otros operadores potenciales de nodos de EOSC se unan a la federación.
- Primer nodo EOSC reconocido a nivel europeo

https://digital-strategy.ec.europa.eu/es/policies/open-science-cloud

European Open Science Cloud (EOSC)

El nodo de la UE EOSC adquirido tiene las siguientes propuestas de valor:

- Ofrece servicios básicos para infraestructuras de investigación científica para federar (es decir, servicios básicos de EOSC, tales como: el inicio de sesión único, los catálogos, el gráfico de conocimientos, el flujo de trabajo de las aplicaciones, el seguimiento, la contabilidad, el servicio de asistencia técnica) y los servicios horizontales comunes de los que pueden beneficiarse los usuarios finales (es decir, los servicios de intercambio de EOSC, tales como: cómputo, contenedores, transferencia de datos, cuadernos, intercambio de archivos, datos de investigación abiertos).

https://digital-strategy.ec.europa.eu/es/policies/open-science-cloud

Reproducibilidad

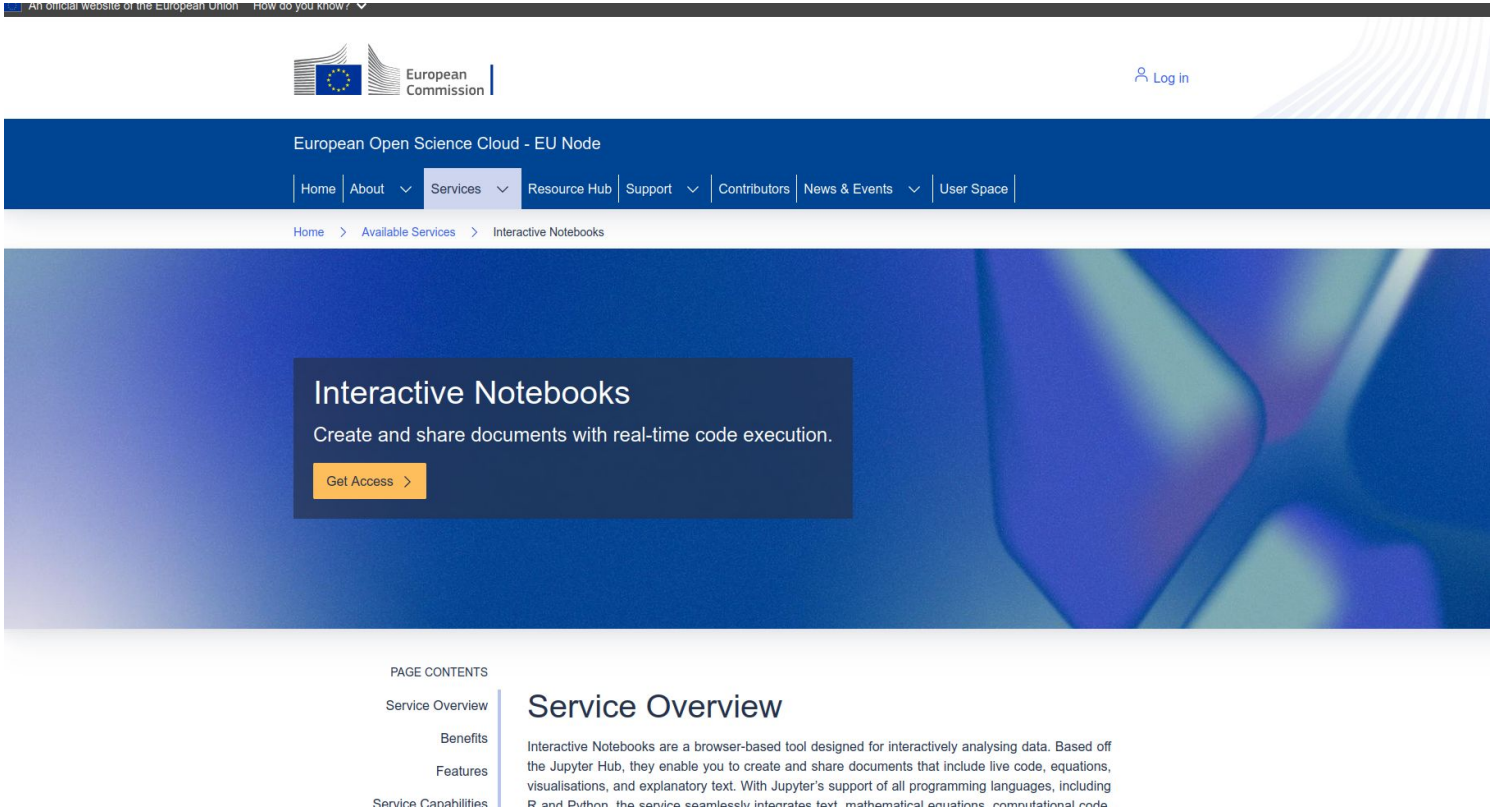
Datos y el código de investigación están disponibles para que otros sean capaces de obtener los mismos resultados alcanzados en sus trabajos.

Jupyter Notebooks ha emergido como una herramiento útil en este sentido para proporcionar **código combinado con documentación textual** ayudando a:

- fomentar la reproducibilidad
- mejorar las capacidades en materia de tecnología de la ciudadanía
- facilitar prototipos demostrando el uso de los datos
- proporcionar una infraestructura de datos donde ejecutar el código

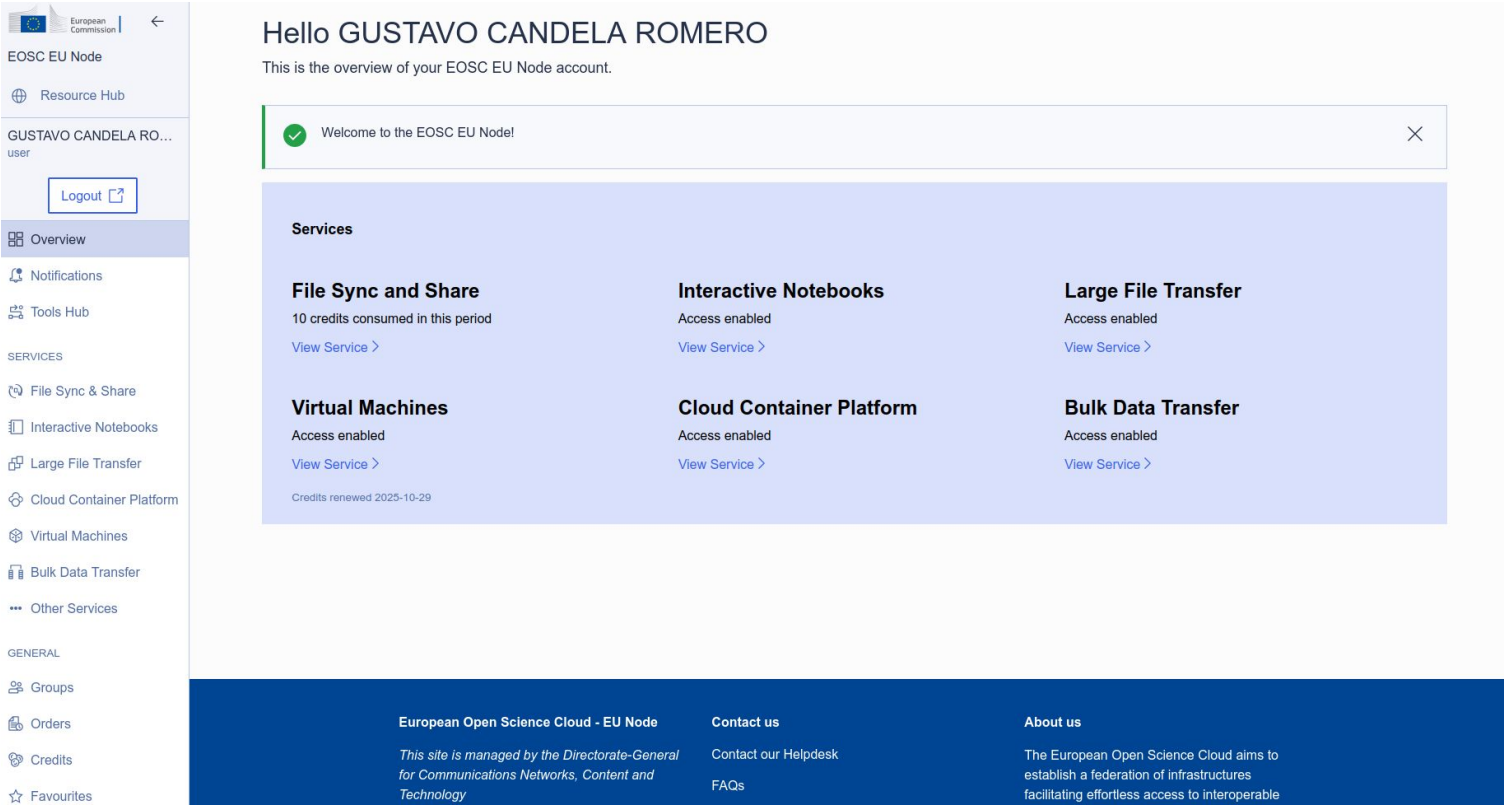
Infraestructuras

https://open-science-cloud.ec.europa.eu/services/interactive-notebooks



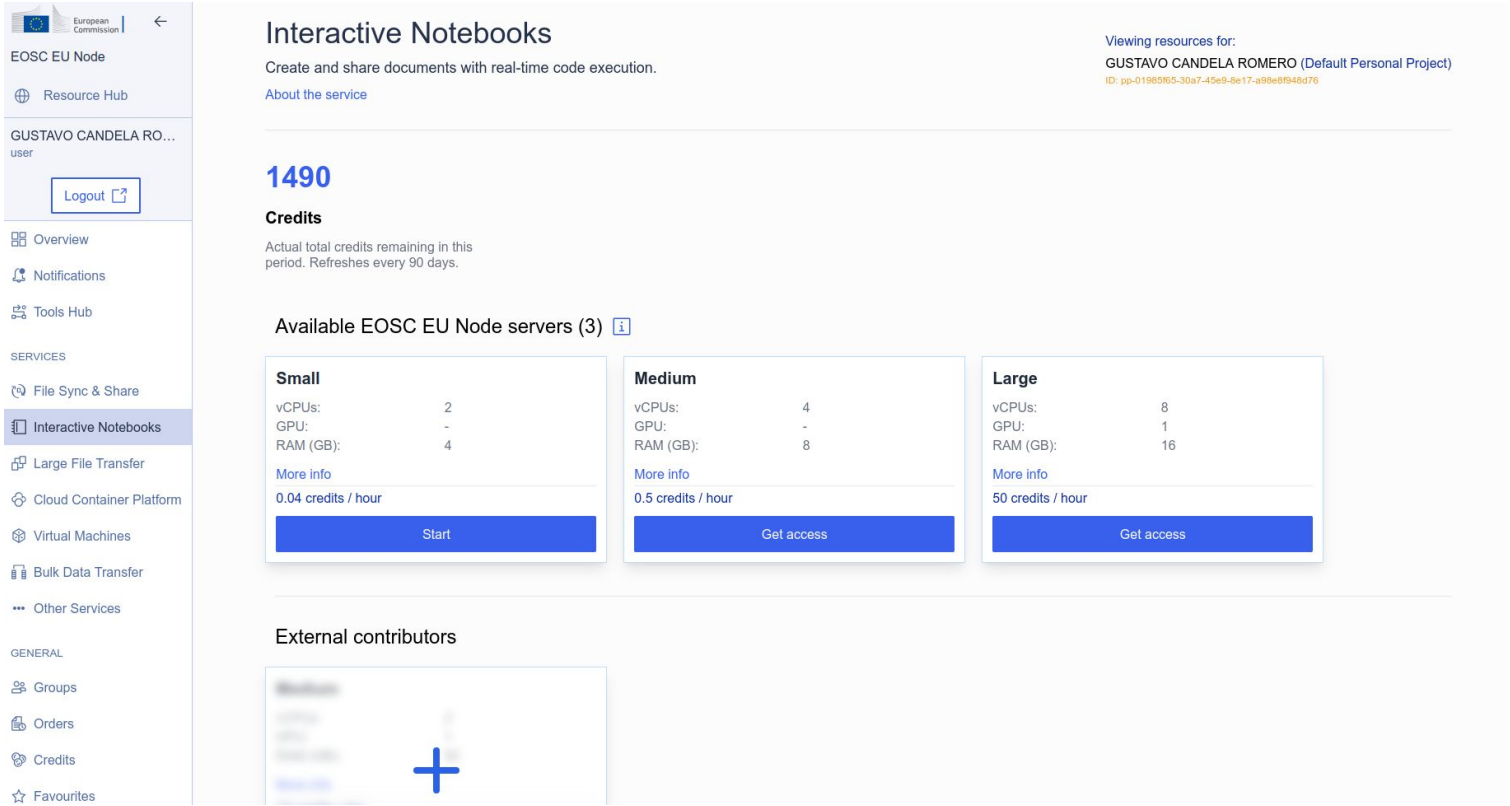
Infraestructuras

European Commission



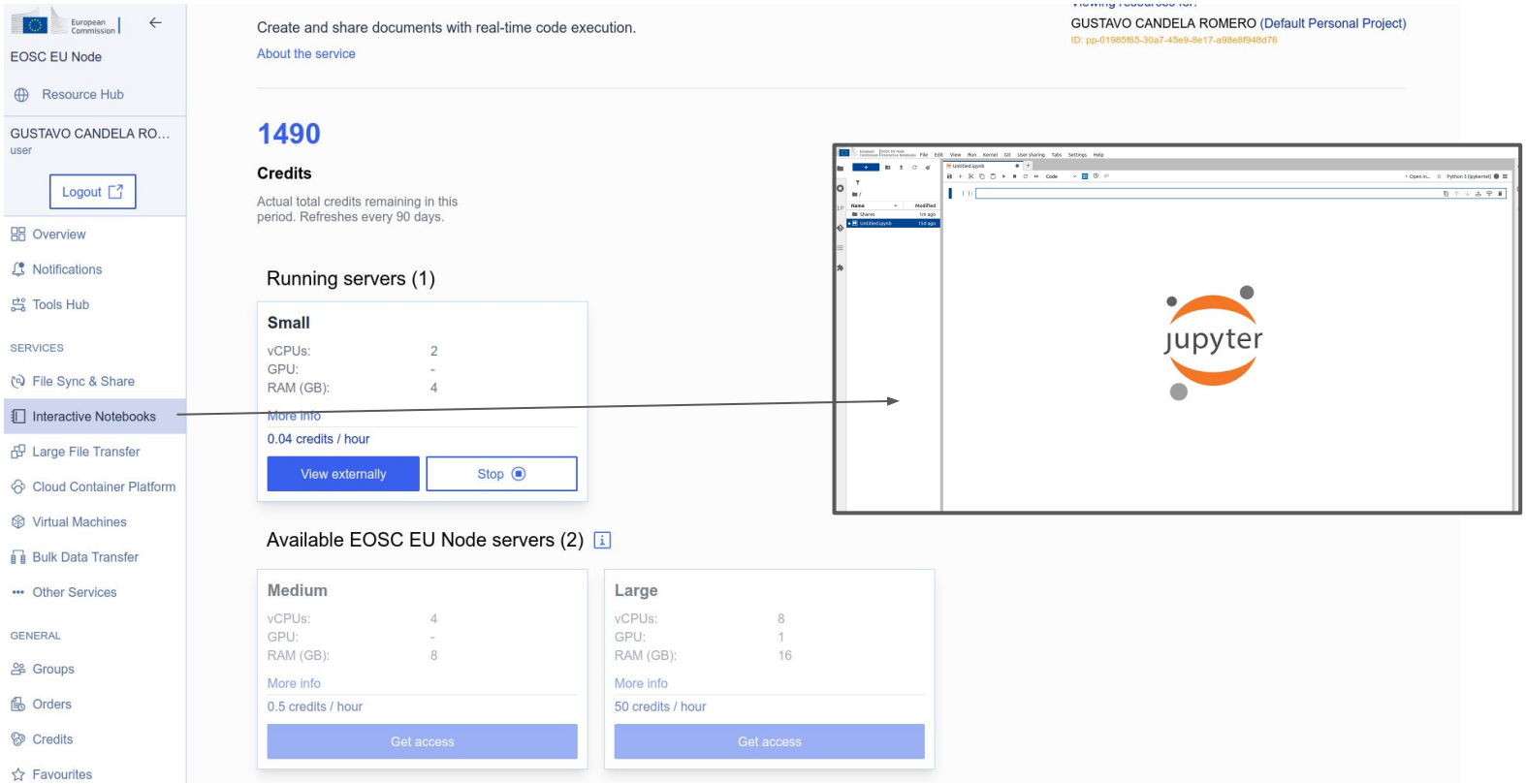
Infraestructuras

European Commission

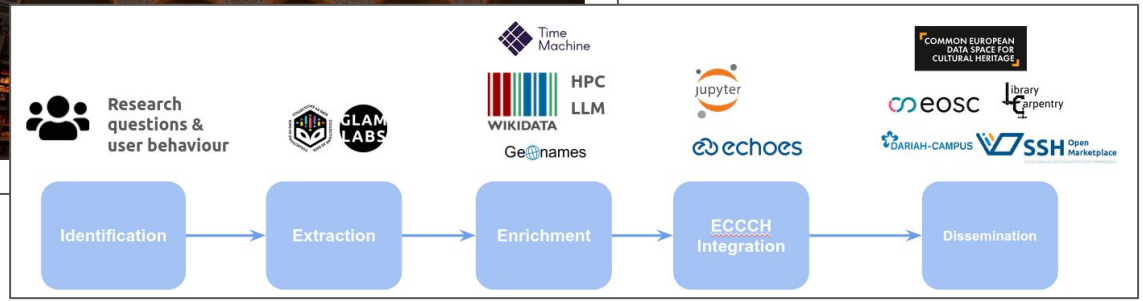


Infraestructuras

European Commission



Cultural Heritage Cloud (ECCCH)



Definiendo casos de uso para el ECCCH

https://doi.org/10.48550/arXiv.2509.08710

Infraestructuras

Jupyter Notebooks Ejemplos

- https://glam-workbench.net/
- Guerra Civil Española y Exilio
- Museos
- Sitios turísticos en Alicante
- Consultas federadas entre varios repositorios
- Visualización de datos
- Visión por computador
- Archivo Web
- Calidad de datos

Infraestructuras

Otros ejemplos

- Apache Airflow: monitorización y creación de workflows
- Apache Kafka: gestión de mensajes
- Apache Spark: procesamiento eficiente de datos a gran escala
- Galaxy Community Hub: creación de workflows ejecutables
- Google Colab
- Hugging Face
- Databricks
- DARIAH-EU y CLARIN

Referencias

- https://dblp.org/pid/160/1759.html
- https://marketplace.sshopencloud.eu/workflow/I3JvP6
- https://mobility-dataspace.eu/
- https://eosc.eu/
- https://www.echoes-eccch.eu/
- https://glamlabs.io/
- https://zenodo.org/
- https://movilidad.lucentia.es/
- https://sophia.lucentia.es/
- https://github.com/hibernator11