

# Procesos ETL

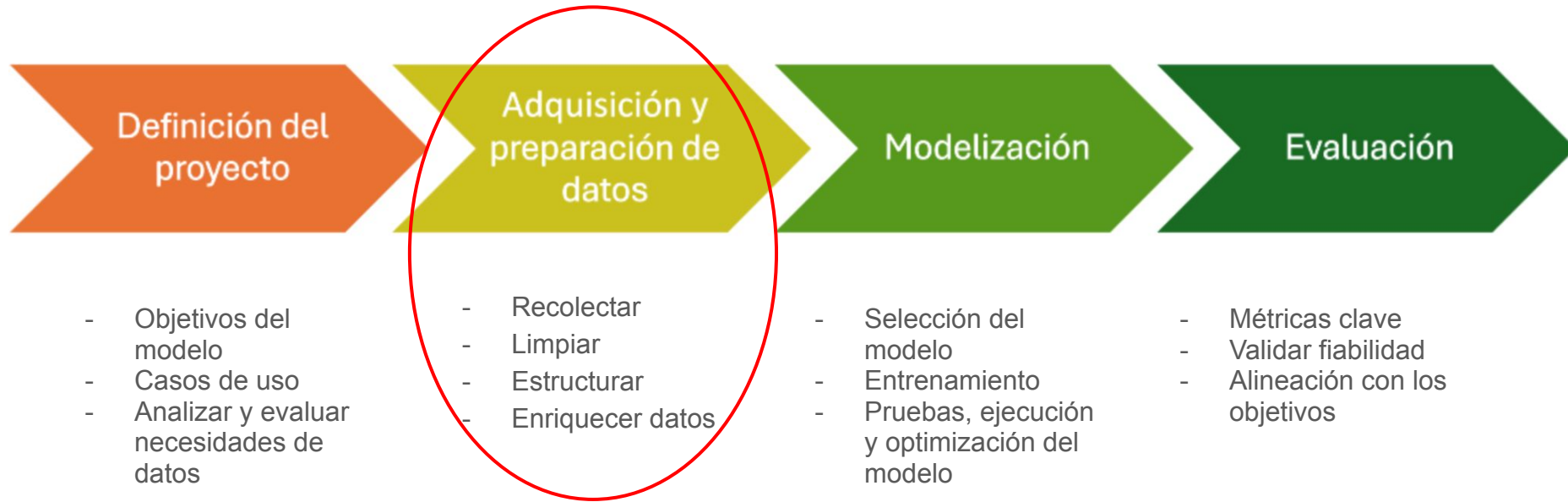
## Tema 5: Introducción a la integración de datos mediante ETLs. Parte I

### Adquisición y preparación de datos

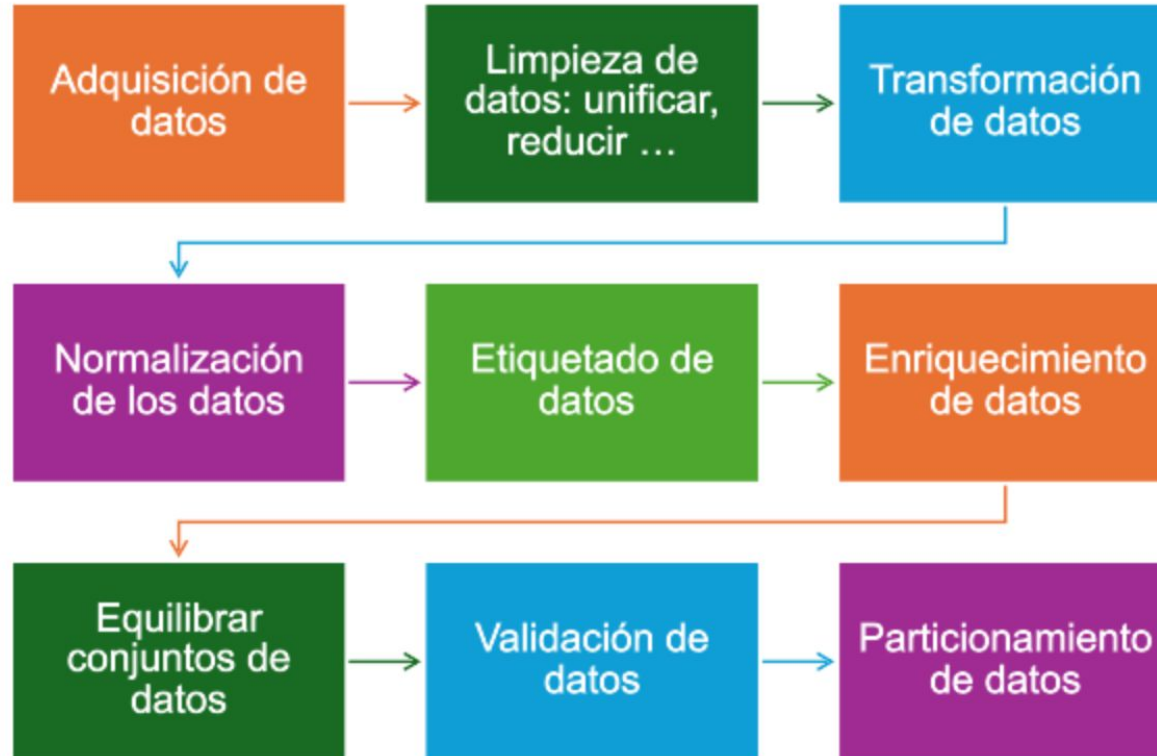


Universitat d'Alacant  
Universidad de Alicante

# 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\_de\_baixa: Mitjana edat homes|població\_resident|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 recollida pneumàtica|0|No visitat|0|No visitat|14|2.2|22|4.2|21|21|4.3|21|2.6|21|1.7|31|32.1|32.4|32.9|47.274|74,6|48

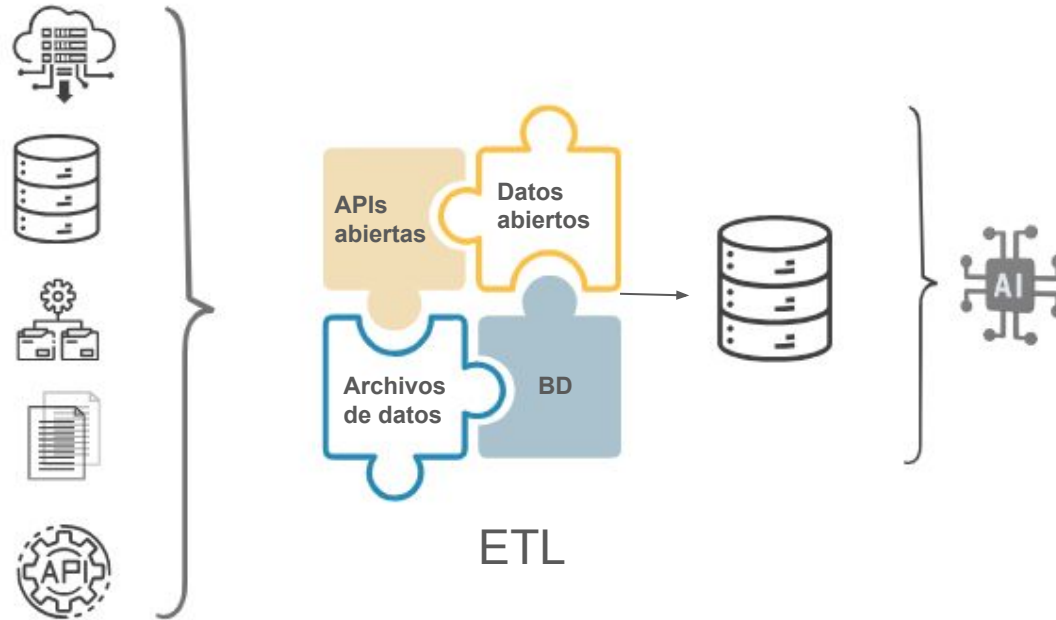
- 1,"4. Sant Pere, Santa Caterina i la Ribera", "32.5", "33.1", "33.1", "34.1"
- 2,"5. el Fort Pienc", "32.6", "33.9", "33.3", "34.9"
- 2,"6. la Sagrada Família", "33.6", "36", "33.2", "35.5"
- 2,"7. la Dreta de l'Eixample", "36.2", "39.6", "34.3", "36.7"
- 2,"8. l'Antiga Esquerra de l'Eixample", "34.6", "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>
  </record>
</collection>
```

# 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.

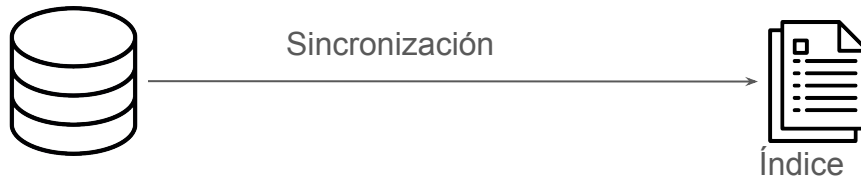
Busca asegurar la **consistencia** y facilitar los datos **en tiempo real** o con mínima latencia.

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**

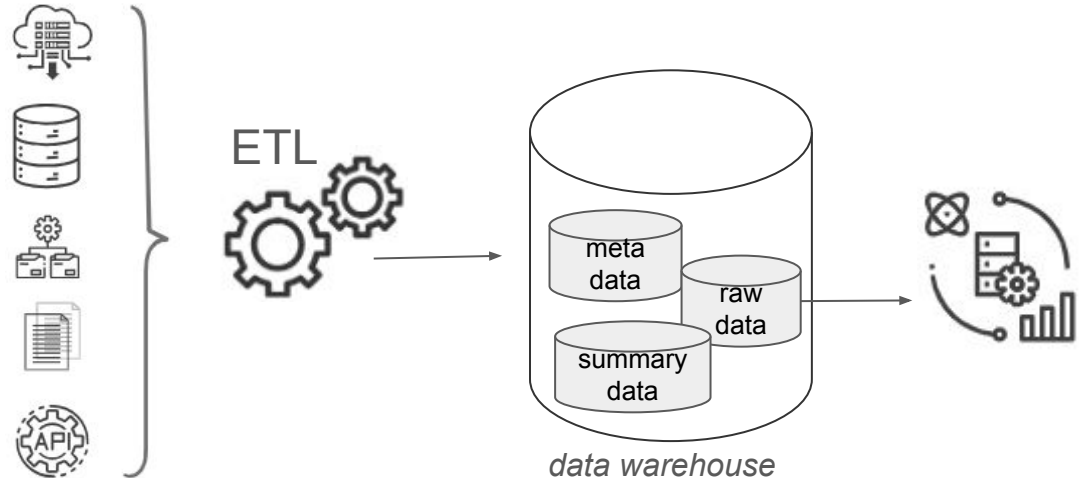
 **elasticsearch**

# 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

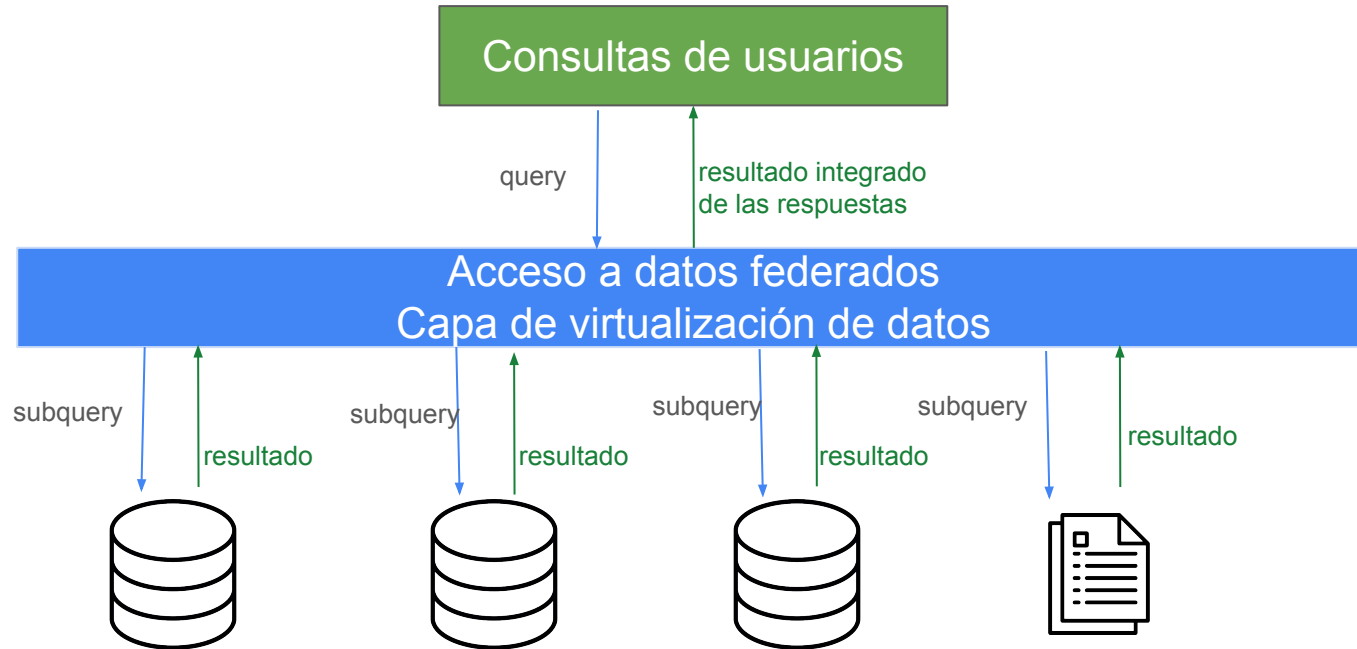
Funcionamiento de una consulta:

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.

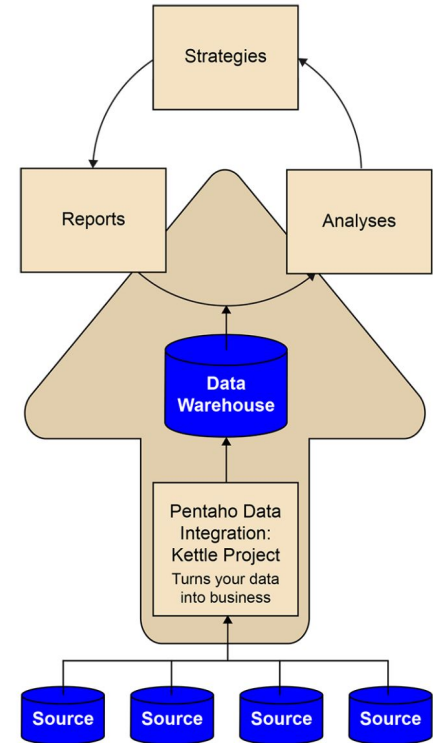
Servicio de integración de datos ETL (extract, transform, load).

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

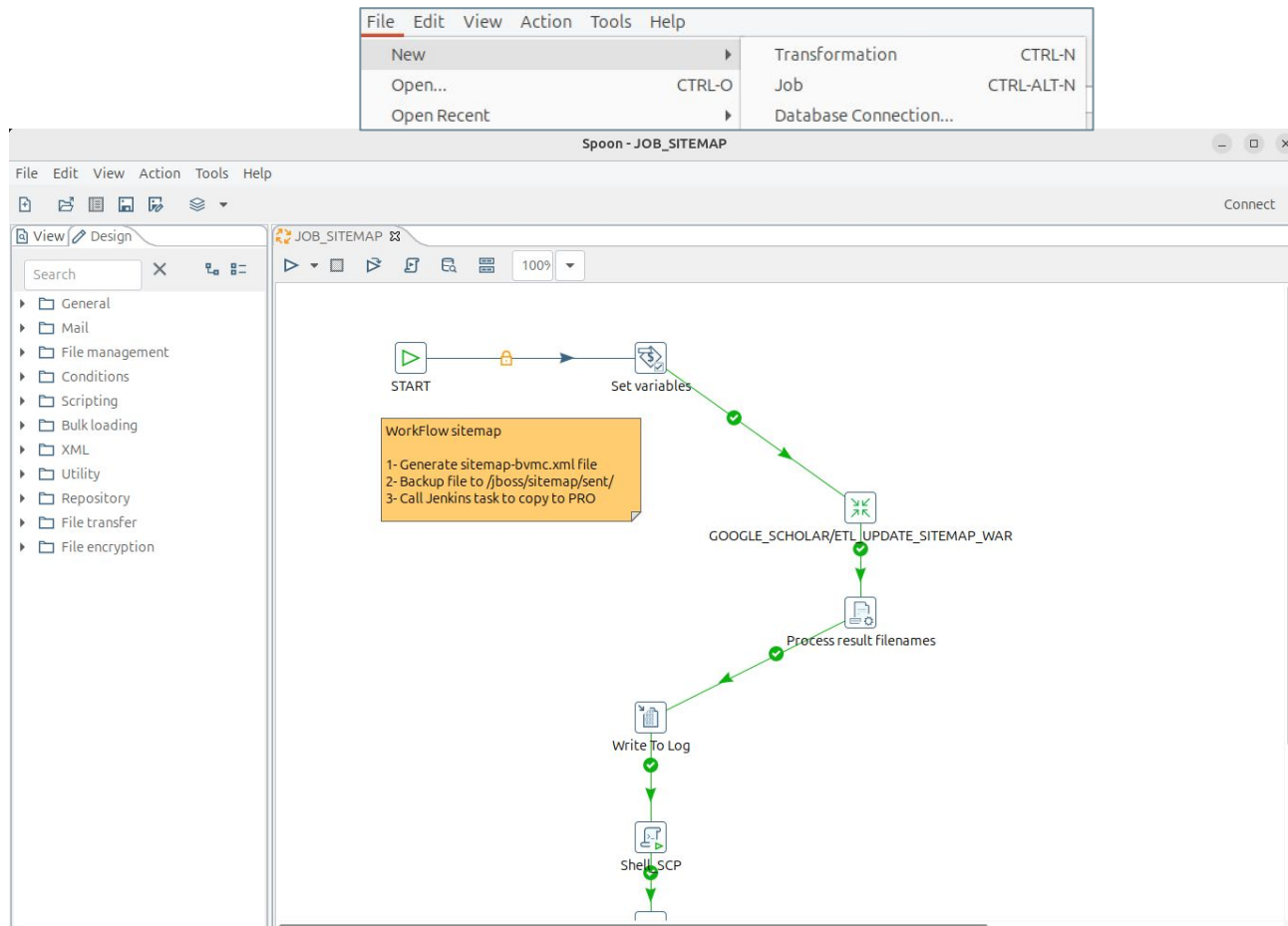


Fuente: Pentaho.



# Pentaho Data Integration

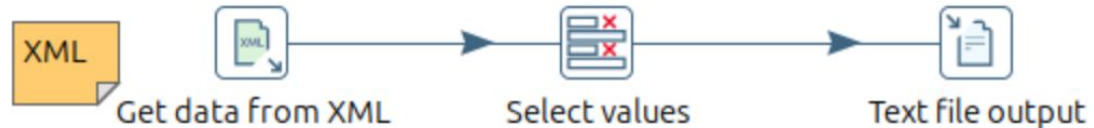
Entorno visual



# Pentaho Data Integration

## *Transformations*

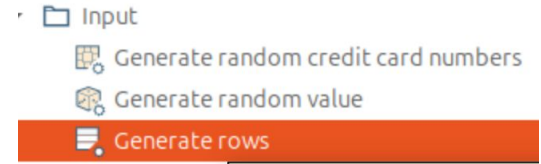
- Proceso encargado de la **manipulación de datos** en el sentido más amplio (extracción, transformación y carga).
- Consta de uno o más **pasos** que realizan diferentes operaciones como leer, filtrar, limpiar o cargar datos en una base de datos.
- Los pasos están conectados mediante **saltos** que 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.

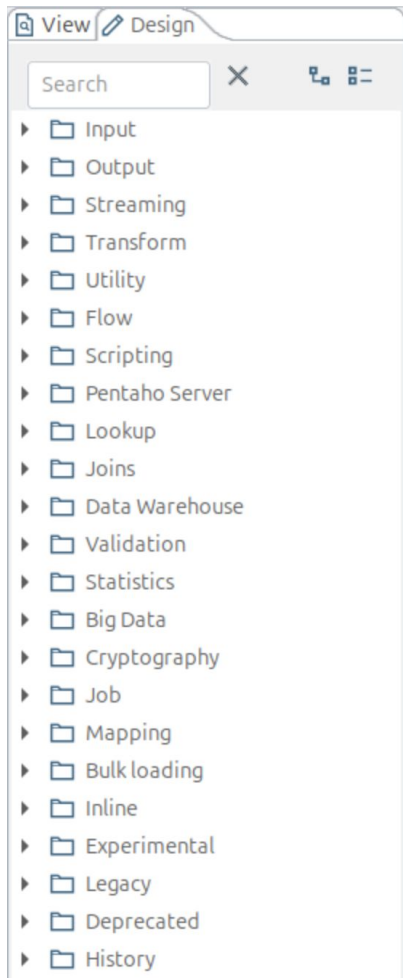


# Pentaho Data Integration

## 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

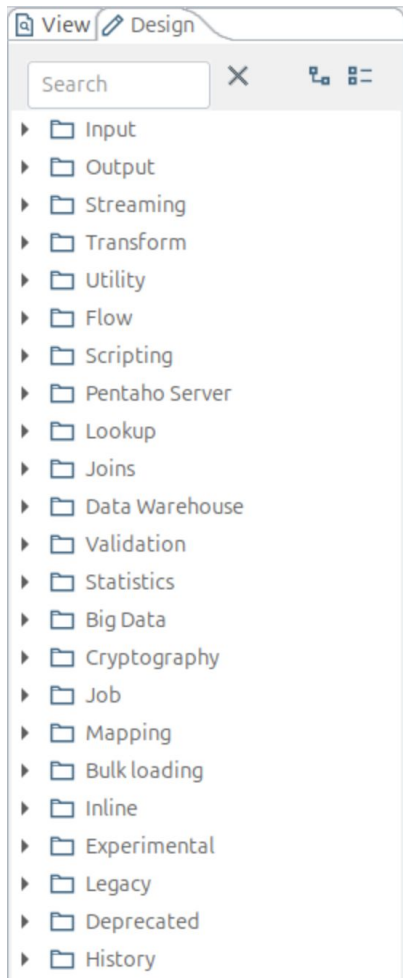
**Utility**, ejecución ssh commands, envío de mensajes syslog, etc

**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.

Transformation properties

Transformation Parameters Logging Dates Dependencies Miscellaneous Monitoring

Nr of rows in rowset 10000

Show a feedback row in transformation ☒

Feedback size 50000

Make the transformation database transactional ☐

Shared objects file

Manage thread priorities? ☒

Transformation engine type Normal

OK SQL Cancel

- No se permiten bucles en transformaciones.

# Pentaho Data Integration

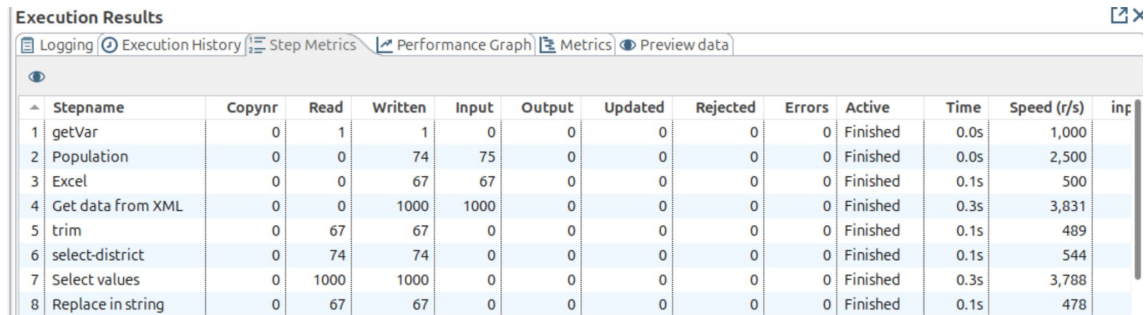
## *Parallelism*

- 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)	inf
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	

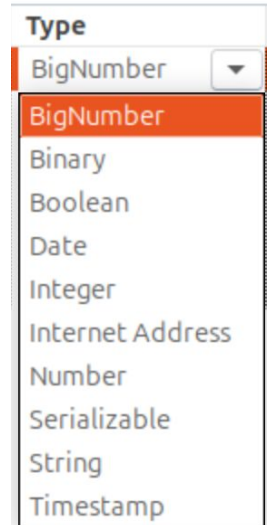
# Pentaho Data Integration

## *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

Select values

Step name

Select & Alter

Remove

Meta-data

Fields to alter the meta-data for :

#	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.		Serializable								
3	Districte d'a				0						
4	Districte d'a				0						
5	Districte de				0						
6	Districte de				0						

Get fields to change

Help

OK

Cancel

# Pentaho Data Integration

**Format**

#,##0.###

#,##0.###

0.00

00000000000000

#.#

#

###,###,###.#

#####.###

#####.###%

Date Locale	Date Time
en_SG	
en_US	
en_ZA	
es	
es_AR	
es_BO	
es_CL	
es_CO	
es_CR	
es_CU	

**Date Time Zone**

Europe/Madrid

Europe/Malta

Europe/Mariehamn

Europe/Minsk

Europe/Monaco

Europe/Moscow

Europe/Nicosia

Europe/Oslo

Europe/Paris

Europe/Podgorica

Europe/Prague

Europe/Riga

Europe/Rome

Europe/Samara

Europe/San\_Marino

Europe/Sarajevo

Europe/Saratov

Europe/Simferopol

Encoding	Decimal
Big5	
ISO-2022-JP	
ISO-2022-JP-2	
ISO-2022-KR	
ISO-8859-1	
ISO-8859-13	
ISO-8859-15	
ISO-8859-16	
ISO-8859-2	
ISO-8859-3	
ISO-8859-4	

# 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 have occurred between them "Y" - uses UTC, meaning it will take into account any time-hop that may have 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
33 KETTLE_DEFAULT_TIMESTAMP_FORMAT		The name of the variable containing an alternative default timestamp format

OKCancel

# 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

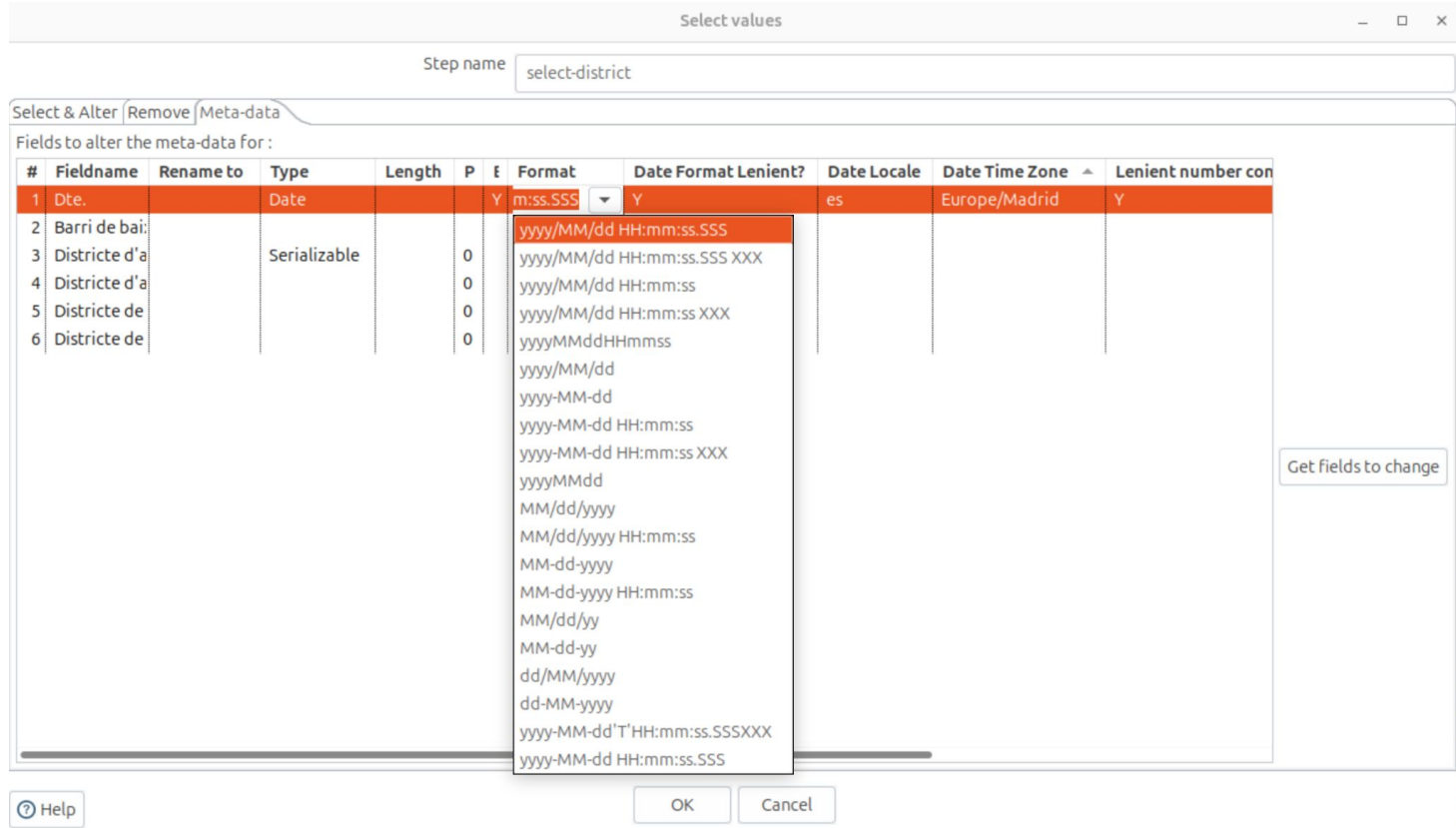
## *Date to String Conversion*

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





# Pentaho Data Integration

## *Numeric to String Conversion*

Los datos numéricos (Number, Integer, BigInteger) 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	<p>This environment variable is used to define which calculation method is to be used by the 'Add a Checksum' step:</p> <ul style="list-style-type: none"><li>• "BYTES" (default) - based on Byte representation of fields; as in versions since 8.1</li><li>• "PENTAHO_STRINGS" - based on Pentaho String representation of fields (applying format masks); as in versions until 7.1</li><li>• "NATIVE_STRINGS" - based on Native String representation of fields; as in version 8.0</li></ul>
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	<p>value:N</p> <p>NULL vs Empty String. If this setting is set to Y, an empty string and null are different. Otherwise they are not</p>
KETTLE_DO_NOT_NORMALIZE_SPACES_ONLY_STRING_TO_EMPTY,	Prevents Kettle from converting spaces only strings to empty string