

# GESTIÓN DE DATOS

Ing. Luis Gabriel Moreno Sandoval MSc MBA PhD(c).  
[morenoluis@javeriana.edu.co](mailto:morenoluis@javeriana.edu.co)

# DOMINIOS DE DATOS

- ¿Cuales son los datos que se encuentran en una organización?

# DOMINIOS DE DATOS



# DOMINIOS DE DATOS

# META DATOS

- “Datos a cerca de los datos”.
- Permiten convertir datos en información.
- Oportunidad es muy importante.
  - Deben estar actualizados.
- Meta datos Técnicos
  - Diagrama ER, Diagramas de Componentes...
- Meta datos de Negocio.
  - Diagramas de Proceso de Negocio, Métricas de Negocio...
- Consistencia entre MD Técnicos y MD de Negocios permite buena comunicación Negocio-IT.

# DATOS MAESTROS

- Entidades principales para el negocio:
  - Clientes.
  - Productos.
  - ¿Más Ejemplos?
- La Confianza es clave.
  - Tienen un impacto muy fuerte.
  - ¿Qué pasa cuando los datos maestros no son confiables?

# DATOS OPERACIONALES

- Datos derivados de las transacciones de negocio.
- Estructurados.
- Alto nivel de detalle.
- Alcance generalmente es Local (LOB).
- Si la confianza en este dominio es buena, facilita los procesos para derivar Datos Analíticos.

# DATOS NO ESTRUCTURADOS

- Tiene un propósito operacional.
  - Se distinguen pues suelen tener requerimientos diferentes.
    - Suelen ser “fotos instantáneas” durante los procesos de negocio.
    - Una vez creados se mantienen estáticos.
  - Contratos, imágenes, notas en un CRM.
- Con el tiempo pueden perder precisión.
  - Ej. La persona que escribió un texto ya no trabaja en la organización.

# DATOS ANALÍTICOS

- Se derivan de los Datos Operacionales, apoyándose en los demás dominios.
- Producir Datos Analíticos requiere normalmente:
  - Integración.
  - Limpieza.
  - Eliminar duplicados.
- Debe tener buenos niveles de precisión, consistencia y relevancia.
- El alcance puede ser tanto local como empresarial.



# MODELO DE REFERENCIA DE INFORMACIÓN




# METADATOS TWITTER

- Campos que definen un usuario.
- Campos que definen un tweet.
- Diccionario de Datos BDs.




# DATOS MAESTROS TWITTER

- Cuentas.
- Relación - *follow*.
- Apps.

 **Luis Gabriel Moreno**  
@gabrielmoreno10


Seguidores

Siguiendo




**Pronto for Business**  
@ProntoBusiness  
Pronto is a communication hub created for the everyday user. It connects people via chat and video, so they can learn faster, work smarter, and communicate seam

Siguiendo




**Forbes Tech** ✓  
@ForbesTech  
Tech news and insights from @Forbes.

Siguiendo




**TechCrunch** ✓  
@TechCrunch  
Technology news and analysis with a focus on founders and startup teams. Got a tip? tips@techcrunch.com

Siguiendo



**MIT Technology Review** ✓  
@techreview  
A media company making technology a greater force for good. Get our journalism: [technologyreview.com/newsletters](https://technologyreview.com/newsletters)

Siguiendo



**MIT Tech Review ES** ✓  
@techreview\_es  
Edición en español de MIT Technology Review, la revista de tecnología más antigua del mundo, publicada por @Opinno [#tecnología](#) [#emprendimiento](#) [#startup](#)

Siguiendo

# DATOS OPERACIONALES (ESTRUCTURADOS)

- Tweets.
- Mensajes.
- Solicitudes.
- Logs.

1	Twitter Own and Retweeted Tweets Table						
2							
3	pageName	createTime	message	link	retweets	favs	replies
4	@realDonaldTrump	09/02/2016 15	#AmericaFirst #ImWithYou	https://twitter.com/r	4267	11671	150
5	@realDonaldTrump	09/02/2016 12	Great new poll Iowa - thank	https://twitter.com/r	6918	17422	271
6	@realDonaldTrump	09/02/2016 08	I visited our Trump Tower c	https://twitter.com/r	5603	21556	1011
7	@realDonaldTrump	09/02/2016 08	People will be very surprise	https://twitter.com/r	7172	21719	420
8	@realDonaldTrump	09/02/2016 08	Just heard that crazy and ve	https://twitter.com/r	4877	16742	983
9	@realDonaldTrump	09/01/2016 18	I will be interviewed by @e	https://twitter.com/r	3611	13986	349
10	@realDonaldTrump	09/01/2016 13	I am promising you a new le	https://twitter.com/r	8680	26267	260
11	@realDonaldTrump	09/01/2016 10	Thank you for having me thi	https://twitter.com/r	5780	18576	156
12	@realDonaldTrump	09/01/2016 06	Poll numbers way up - maki	https://twitter.com/r	9351	35537	446
13	@realDonaldTrump	09/01/2016 06	Thank you to @foxandfrien	https://twitter.com/r	6641	26977	220
14	@realDonaldTrump	09/01/2016 06	Mexico will pay for the wall	https://twitter.com/r	27781	53159	1793
15	@realDonaldTrump	09/01/2016 01	Under a Trump administrati	https://twitter.com/r	7991	21867	377
16	@realDonaldTrump	09/01/2016 01	Hillary Clinton doesn't have	https://twitter.com/r	6338	18987	292

```
"created_at" : "Mon Oct 12 16:27:42 +0000 2020",
"id" : NumberLong(1315690618018988033),
"id_str" : "1315690618018988033",
"text" : "Que hpta maricada en mi casa con este tema",
"source" : "<a href='\"http://twitter.com/download/android\"' rel='\"nofollow\"'>Twitter for Android</a>",
"truncated" : false,
"in_reply_to_status_id" : null,
"in_reply_to_status_id_str" : null,
"in_reply_to_user_id" : null,
"in_reply_to_user_id_str" : null,
"in_reply_to_screen_name" : null,
"user" : {
  "id" : 344242778,
  "id_str" : "344242778",
  "name" : "Jennifer A.",
  "screen_name" : "JenniferCampoV",
  "location" : "Guadalajara de Buga, Colombia",
  "url" : null,
  "description" : "♥MACARENA♥",
  "translator_type" : "none",
  "protected" : false,
  "verified" : false,
  "followers_count" : 624,
  "friends_count" : 350,
  "listed_count" : 1,
  "favourites_count" : 8629,
  "statuses_count" : 17080,
  "created_at" : "Thu Jul 28 19:19:14 +0000 2011",
  "utc_offset" : null,
  "time_zone" : null,
  "geo_enabled" : true,
  "lang" : null,
  "contributors_enabled" : false,
  "is_translator" : false,
  "profile_background_color" : "FF6699",
  "profile_background_image_url" : "http://abs.twimg.com/images/themes/themell/bg.gif",
  "profile_background_image_url_https" : "https://abs.twimg.com/images/themes/themell/bg.gif",
  "profile_background_tile" : true,
```

# DATOS NO ESTRUCTURADOS

- Texto de tweets.
- Mensajes.
- Fotos.
- Videos.



# DATOS ANALÍTICOS


- Tendencias.
- Top hashtags.
- Estadísticas.
- #RTs.
- #Favs.

Para ti **Tendencias** COVID-19 Noticias Deportes Entretenimiento

## Tendencias de Colombia

1 • Liga de Campeones de la UEFA • Tendencia ...  
**Barcelona**

Liga de Campeones de la UEFA  
UCL: Barcelona cae en casa ante el Bayern Múnich ⚽




256 mil Tweets

2 • Tendencias ...  
**#TambienMeEngañaron**  
1.062 Tweets

3 • Tendencias ...  
**iPhone 13**

Tecnología  
Apple presenta el iPhone 13, el Apple Watch Series 7 y nuevos iPad 🍏

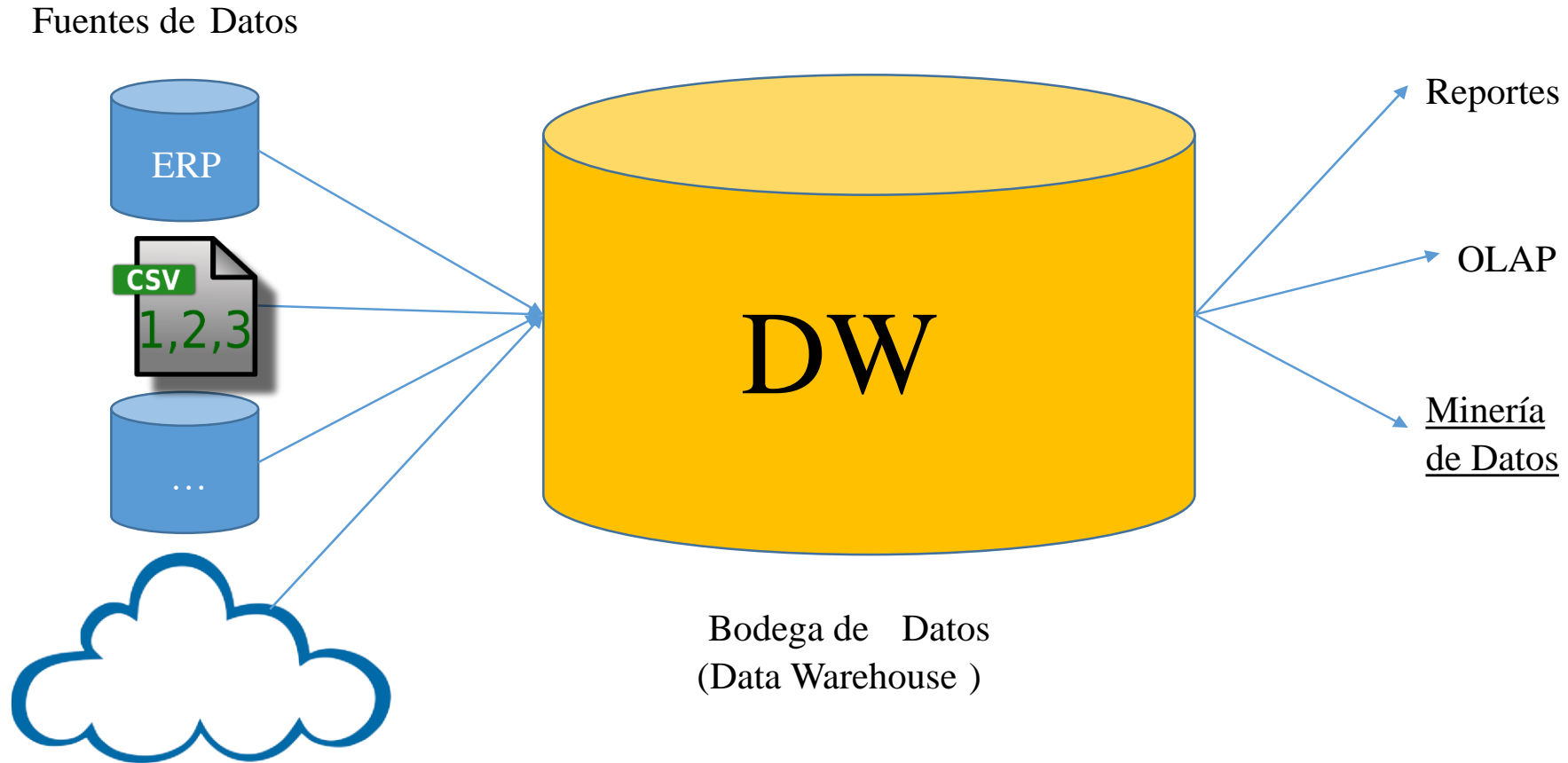


79,1 mil Tweets

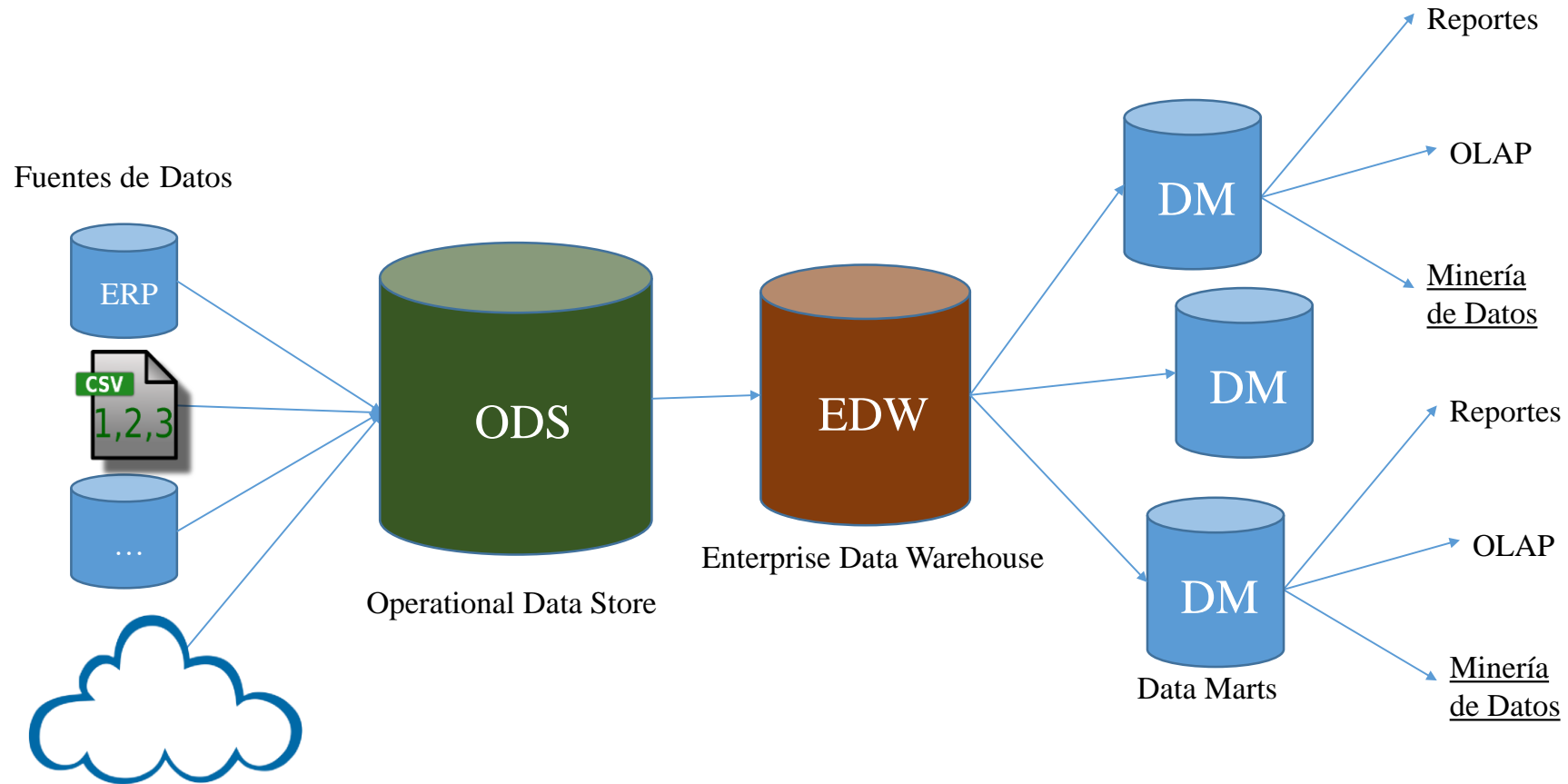
4 • Tendencias ...  
**#PigCoin**  
2.158 Tweets

5 • Tendencias ...  
**Tienen 24**  
1.160 Tweets

# MAPA GENERAL DE DATOS ANALÍTICOS



# VERSIÓN COMPLETA DE DW





# TIPOS DE BODEGAS DE DATOS

- Operational Data Store.
  - Datos limpios.
  - Única fuente de verdad.
  - Relativamente Normalizados.
  - Alta cantidad de detalles.
  - Corto rango de tiempo.
- Enterprise Data Warehouse.
  - Multidimensional (Dimensión y hechos).
    - No está normalizado.
  - Nivel de detalle menor a ODS.
    - Algunas agregaciones/resúmenes.
  - Rango de tiempo mayor.

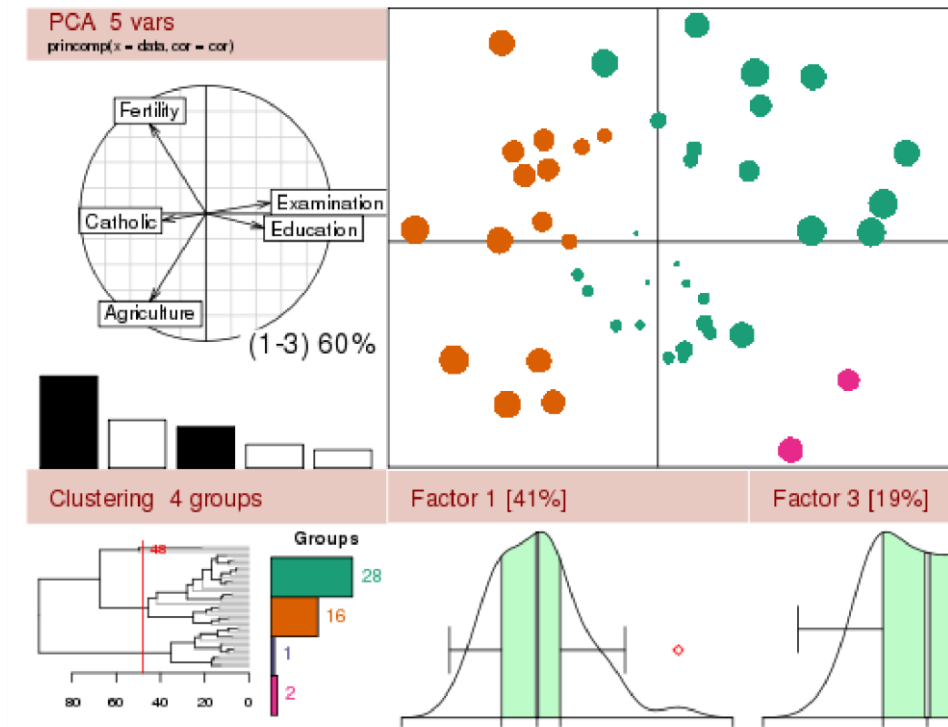
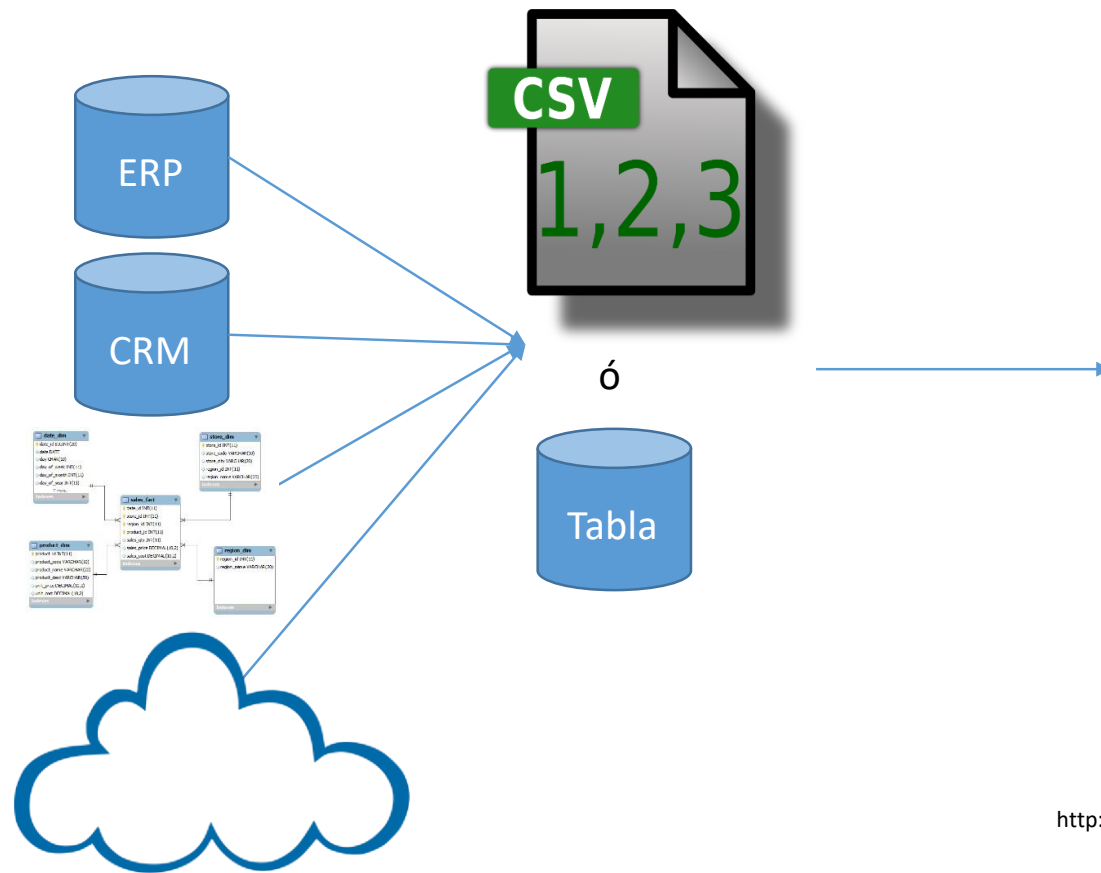
# TIPOS DE BODEGAS DE DATOS

- Data Mart.
  - Multidimensional.
  - Orientado a un tema.
  - Agregaciones y resúmenes.

<b>TIPO</b>	<b>DETALLE</b>	<b>ALCANCE</b>	<b>HISTORIAL</b>	<b>USO</b>
<b>ODS</b>	Máximo nivel de detalle.	Empresa.	Corto.	Táctico, análisis de la operación en el día a día.
<b>EDW</b>	Detalle y agregaciones.	Empresa.	Amplio.	Táctico y Estratégico.
<b>DM</b>	Agregaciones y Poco detalle.	Tema específico.	Amplio.	Datos con alto valor para un grupo de personas, unidad de negocio...

# TIPOS DE BODEGAS DE DATOS

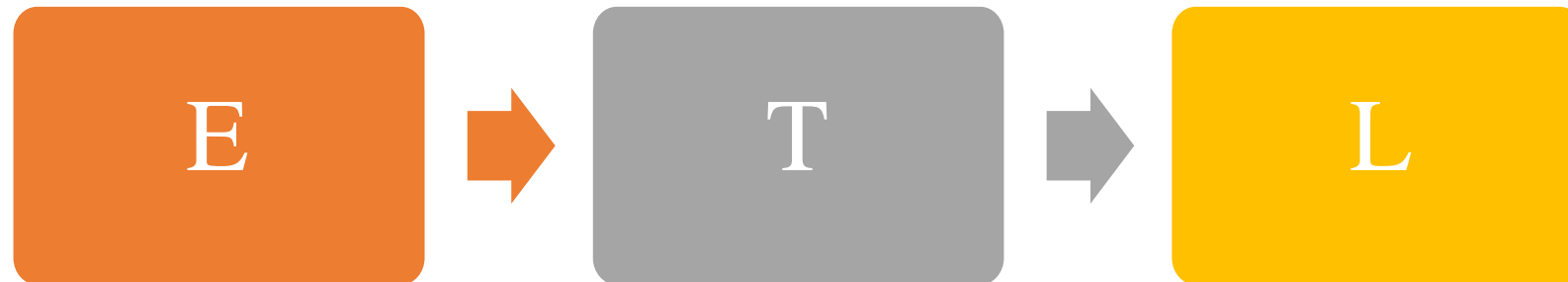
# VERSIÓN COMPACTA DE DATOS ANALÍTICOS



<http://www.datajournalismtools.net/post/65217067609/> - a free software environment for statistical

# ETL

- Extracción.
  - Conectarse a una fuente de datos.
  - Seleccionar y leer los datos relevantes.
- Transformación.
  - Hacer operaciones sobre los datos.
- Carga.
  - Guardar el resultado en un destino.



# EXTRACCIÓN – FUENTES DE DATOS MÁS COMUNES

- Archivos CSV (comma-separated values).
  - Archivos que contienen datos tabulares.
- DBMS (Database Management System).
  - Bases de datos tradicionales: Oracle, SQL Server, MySQL ...
  - Lenguaje de consultas SQL.
- JSON (JavaScript Object Notation).
- XML (eXtensible Markup Language).
- APIs RESTful.
  - Fuentes de datos que se consumen por HTTP.

# CSV (COMMA-SEPARATED VALUES)

- Es un archivo en el que los datos están organizados en FILAS y los atributos se separan entre sí por comas (o punto y coma).
- La primera fila suele ser el encabezado.
- Pueden ser visualizados en herramientas como Notepad++, LibreOffice o Excel (No recomendado).

# CSV - EJEMPLO

- Notepad++

```
1 codigo_departamento,nombre_departamento,codigo_municipio,nombre_municipio,codigo_postal,tipo,barrios_contenidos_en_el_codigo_postal
2 5,ANTIOQUIA,05001,MEDELLIN,050001,Urbano,Andalucía- La Francia- La Frontera- La Isla- Pablo VI- Playón de Los Comuneros- Villa del S
3 5,ANTIOQUIA,05001,MEDELLIN,050002,Urbano,Aldea Pablo VI- Carpinelo- El Compromiso- La Avanzada- La Esperanza No.2- Popular- Santo Dc
4 5,ANTIOQUIA,05001,MEDELLIN,050003,Urbano,Granizal- La Rosa- Moscú No.1- Moscú No.2- San Pablo- Santa Cruz- Villa Guadalupe
5 5,ANTIOQUIA,05001,MEDELLIN,050004,Urbano,Aranjuez- Berlin- Bermejál-Los Alamos- La piñuela- Palermo- San Isidro
```

- LibreOffice

	A	B	C	D	E	F	
1	codigo_departamento	nombre_departamento	codigo_municipio	nombre_municipio	codigo_postal	tipo	barrios_contenidos_en_el_codigo_postal
2		5 ANTIOQUIA	05001	MEDELLIN	050001	Urbano	Andalucía- La Francia- La Frontera- La Isla- Pablo VI- Playón de L
3		5 ANTIOQUIA	05001	MEDELLIN	050002	Urbano	Aldea Pablo VI- Carpinelo- El Compromiso- La Avanzada- La Espe
4		5 ANTIOQUIA	05001	MEDELLIN	050003	Urbano	Granizal- La Rosa- Moscú No.1- Moscú No.2- San Pablo- Santa C
5		5 ANTIOQUIA	05001	MEDELLIN	050004	Urbano	Aranjuez- Berlin- Bermejál-Los Alamos- La piñuela- Palermo- San
6		5 ANTIOQUIA	05001	MEDELLIN	050005	Urbano	Campo Valdés No.2- La Salle- Las Granjas- María Cano-Carambol
7		5 ANTIOQUIA	05001	MEDELLIN	050006	Urbano	Brasilia- Campo Valdés No.1- Las Esmeraldas- Miranda- Moravia



# OPCIONES CSV

- Separador.
  - “,” ó “;”
  - TSV – separado por tabs (“\t”).
- Primera fila con cabeceras.
  - Puede tenerlas o puede no tenerlas.
- Comillas para texto.
  - Opcionales. Recomendables aunque algunas herramientas no los soportan.
  - Permiten diferenciar un texto de un número.
    - E.g. Teléfono.
- Evita que al leer el archivo se eliminen ceros a la izquierda
  - “018000123456” -> 18.000.123.456 ❌

# EJEMPLO

Separador=";"

Con cabecera

Texto encerrado en comillas dobles

```
1 "mes";presencial;virtual;telefonico;escrito;total_mes
2 "jul-12"; 39414,0; 465; 4997,0; 4218,0; 49094,0
3 "ago-12"; 39858,0; 623; 5791,0; 3663,0; 49935,0
4 "sep-12"; 37490,0; 1519; 4974,0; 4485,0; 48468,0
5 "oct-12"; 39702,0; 1897; 6438,0; 5603,0; 53640,0
6 "nov-12"; 38753,0; 1171; 6486,0; 4772,0; 51182,0
7 "dic-12"; 27051,0; 1210; 6641,0; 5372,0; 40274,0
8 "ene-13"; 38097,0; 1794; 8444,0; 8946,0; 57281,0
```

Separador=","

Sin cabecera

Texto sin encerrar en comillas

```
1 jul-12, 39414,0, 465, 4997,0, 4218,0, 49094,0
2 ago-12, 39858,0, 623, 5791,0, 3663,0, 49935,0
3 sep-12, 37490,0, 1519, 4974,0, 4485,0, 48468,0
4 oct-12, 39702,0, 1897, 6438,0, 5603,0, 53640,0
5 nov-12, 38753,0, 1171, 6486,0, 4772,0, 51182,0
6 dic-12, 27051,0, 1210, 6641,0, 5372,0, 40274,0
7 ene-13, 38097,0, 1794, 8444,0, 8946,0, 57281,0
```

!!!!!!!!!!!!

# CSV - EJERCICIO

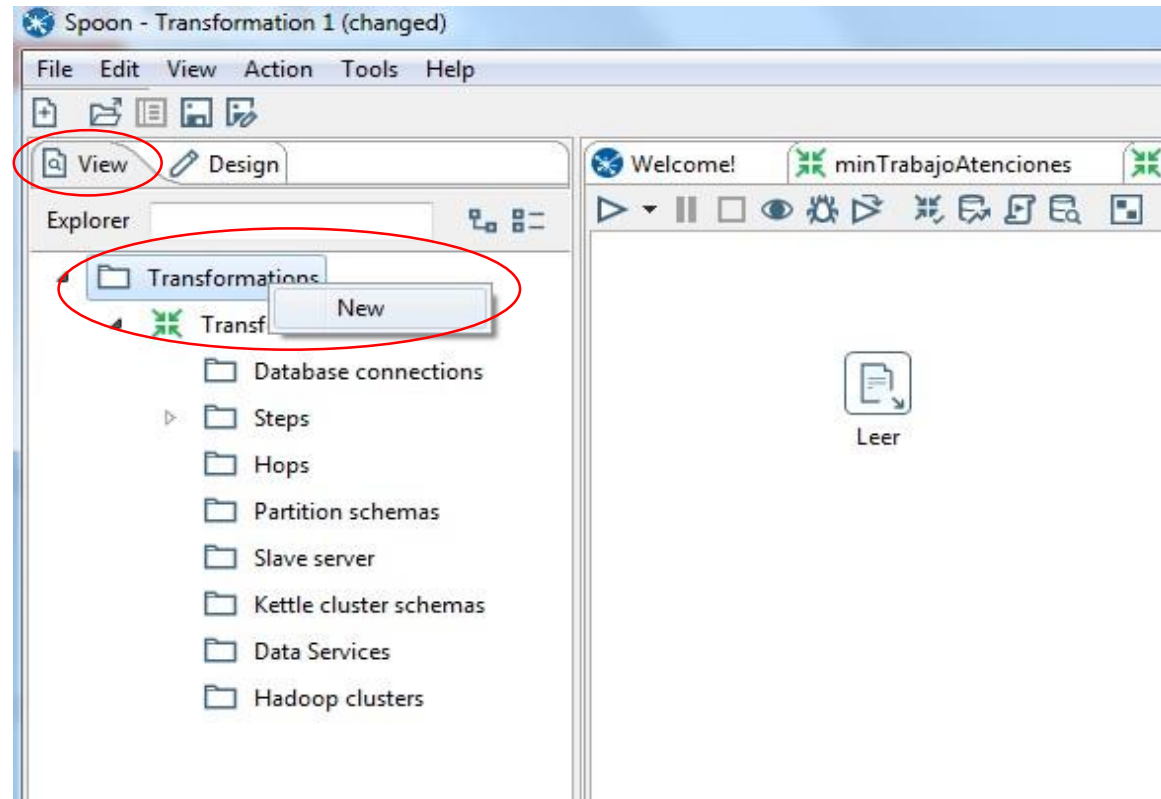
- Abrir el archive `codigosPostalesMedellin.csv`
  - Click derecho - > Abrir con -> Bloc de Notas (Notepad++).
- Explorar los datos.
- Identificar.
  - ¿Separador?
  - ¿Tiene cabecera?
  - ¿Texto en comillas?
- ¿Qué tipo de datos tiene cada columna?
  - Entero, Texto, Fecha, Real.

# LEER CSV CON PENTAHO (SPOON – KETTLE)

- ¿Pentaho?
  - Suite de Inteligencia de Negocios.
  - Versión libre y versión empresarial.
  - Múltiples componentes.
    - El componente libre de ETL se llama Kettle o Spoon.
    - <http://community.pentaho.com/projects/data-integration/>

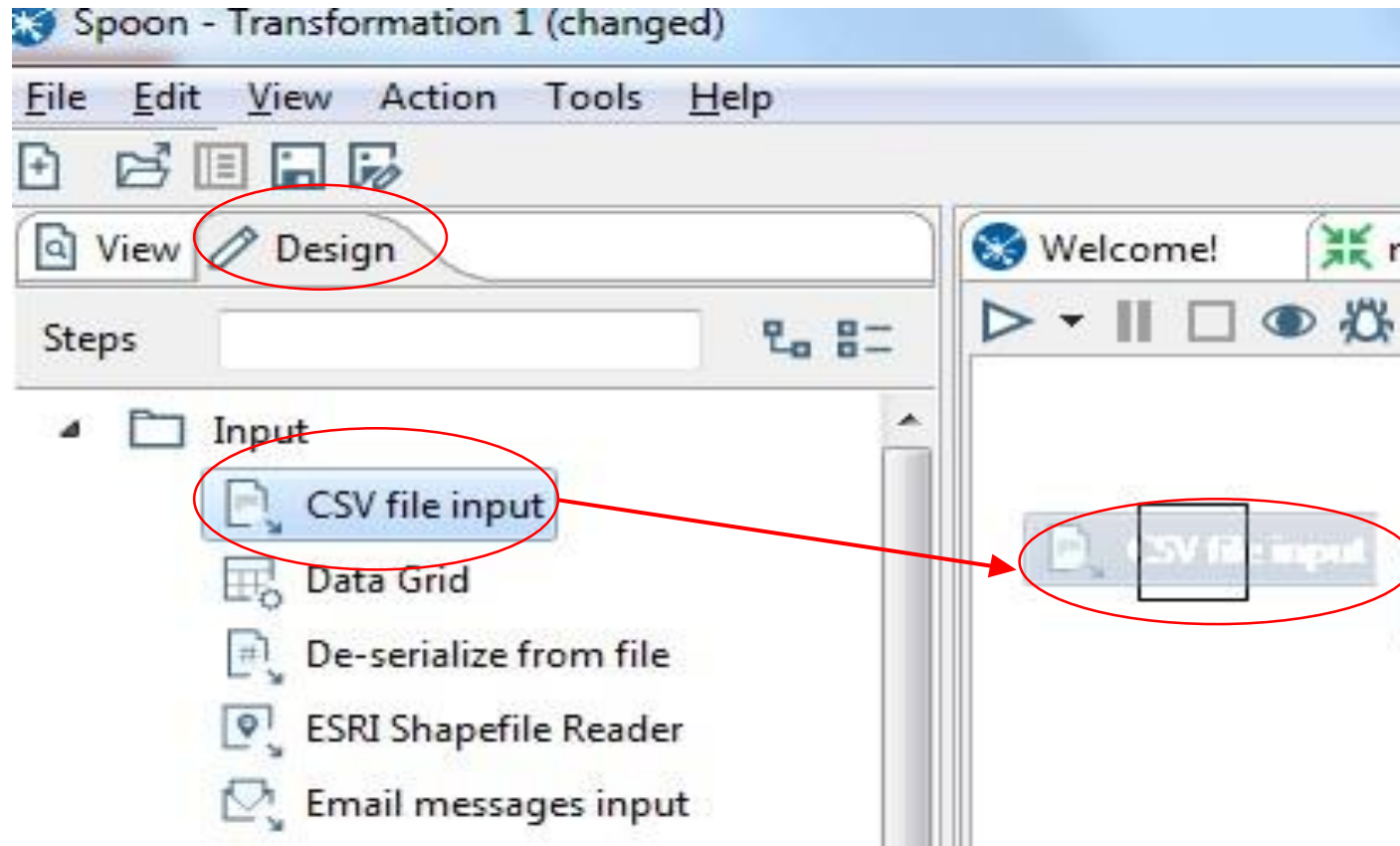
# LEER CSV CON PENTAHO

1. C:/data-integration/Spoon.bat
2. Crear una transformación.



# LEER CSV CON PENTAHO

3. Crear un input de tipo CSV File Input.



# CONFIGURAR EXTRACCIÓN

CSV Input

Step name: Leer CSV

Filename: C:\Users\asierra\Google Drive\javeriana\Gestion de Datos\Ejemplo\codigosPostalesMedellin.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? ☐

File encoding: UTF-8

## TRAER INFORMACIÓN DE LOS CAMPOS

- Pentaho analiza las primeras filas del archivo para tratar de descubrir que tipos de datos hay.

[illegible]



# TRAER INFORMACIÓN DE LOS CAMPOS

- Resultado del análisis.

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	codigo_departamento	Integer	#	15	0	\$	,	.	none
2	nombre_departamento	String		9		\$	,	.	none
3	codigo_municipio	Integer	#	15	0	\$	,	.	none
4	nombre_municipio	String		8		\$	,	.	none
5	codigo_postal	Integer	#	15	0	\$	,	.	none
6	tipo	String		6		\$	,	.	none
7	barrios_contenidos_en_el_codigo_postal	String		254		\$	,	.	none

Help

OK

Get Fields

Preview

Cancel

- Pre visualizar.

Examine preview data

Rows of step: Leer CSV (41 rows)

#	codigo_departamento	nombre_departamento	codigo_municipio	nombre_municipio	codigo_postal	tipo	barrios_contenidos_en_el_codigo_postal
1	5	ANTIOQUIA	5001	MEDELLIN	50001	Urbano	Andalucía- La Francia- La Frontera- La Isla- Pablo VI- Playón de
2	5	ANTIOQUIA	5001	MEDELLIN	50002	Urbano	Aldea Pablo VI- Carpinelo- El Compromiso- La Avanzada- La Es
3	5	ANTIOQUIA	5001	MEDELLIN	50003	Urbano	Granizal- La Rosa- Moscú No.1- Moscú No.2- San Pablo- Santa
4	5	ANTIOQUIA	5001	MEDELLIN	50004	Urbano	Aranjuez- Berlin- Bermejál-Los Alamos- La piñuela- Palermo- Si
5	5	ANTIOQUIA	5001	MEDELLIN	50005	Urbano	Campo Valdés No.2- La Salle- Las Granjas- María Cano-Caramb
6	5	ANTIOQUIA	5001	MEDELLIN	50006	Urbano	Brasilia- Campo Valdés No.1- Las Esmeraldas- Miranda- Moravi
7	5	ANTIOQUIA	5001	MEDELLIN	50007	Rural	Sin Informacion de Barrios
8	5	ANTIOQUIA	5001	MEDELLIN	50010	Urbano	El Chagualo- Estación Villa- Hospital San Vicente de Paúl- Jardí
9	5	ANTIOQUIA	5001	MEDELLIN	50011	Urbano	Batallón Girardot- El Pomar- El Raizal- La Cruz- Manrique Centri
10	5	ANTIOQUIA	5001	MEDELLIN	50012	Urbano	Boston- La Candelaria- La Mansión- Los Ángeles- Prado- San M
11	5	ANTIOQUIA	5001	MEDELLIN	50013	Urbano	Barrio Caycedo- El Pinal- Enciso- La Ladera- La Libertad- LLana
12	5	ANTIOQUIA	5001	MEDELLIN	50014	Urbano	Alejandro Echavarría- Barrios de Jesús- Juan Pablo II- La Sierra- I
13	5	ANTIOQUIA	5001	MEDELLIN	50015	Urbano	Calle Nueva- Centro Administrativo- Corazón de Jesús- Guayaq
14	5	ANTIOQUIA	5001	MEDELLIN	50016	Urbano	Asomadera No.1- Barrio Colón- Bomboná No.1- Buenos Aires-
15	5	ANTIOQUIA	5001	MEDELLIN	50017	Rural	Sin Informacion de Barrios
16	5	ANTIOQUIA	5001	MEDELLIN	50018	Rural	Sin Informacion de Barrios
17	5	ANTIOQUIA	5001	MEDELLIN	50020	Urbano	Bomboná No.2- Cataluña- La Milagrosa- Loreto- Los Cerros El V
18	5	ANTIOQUIA	5001	MEDELLIN	50021	Urbano	Altos del Poblado- Asomadera No.2- Asomadera No.3- Barrio C

¿Está bien?

# AJUSTAR TIPOS DE DATOS

Filename: C:\Users\asierra\Google Drive\javeriana\Gestion de Datos\Ejemplo\codigosPostalesMedellin.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? ☐

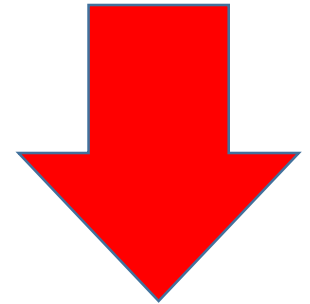
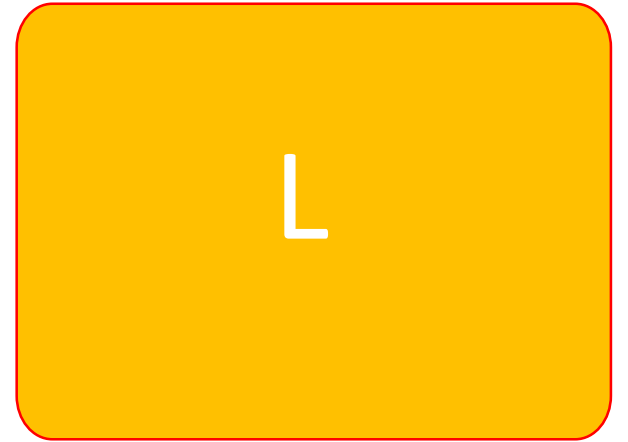
New line possible in fields? ☐

File encoding: UTF-8

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	codigo_departamento	Integer	#	15	0	\$	,	.	none
2	nombre_departamento	String		9		\$	,	.	none
3	codigo_municipio	String	#	15	0	\$	,	.	none
4	nombre_municipio	String		8		\$	,	.	none
5	codigo_postal	String	#	15	0	\$	,	.	none
6	tipo	String		6		\$	,	.	none
7	barrios_contenidos_en_el_codigo_postal	String		254		\$	,	.	none

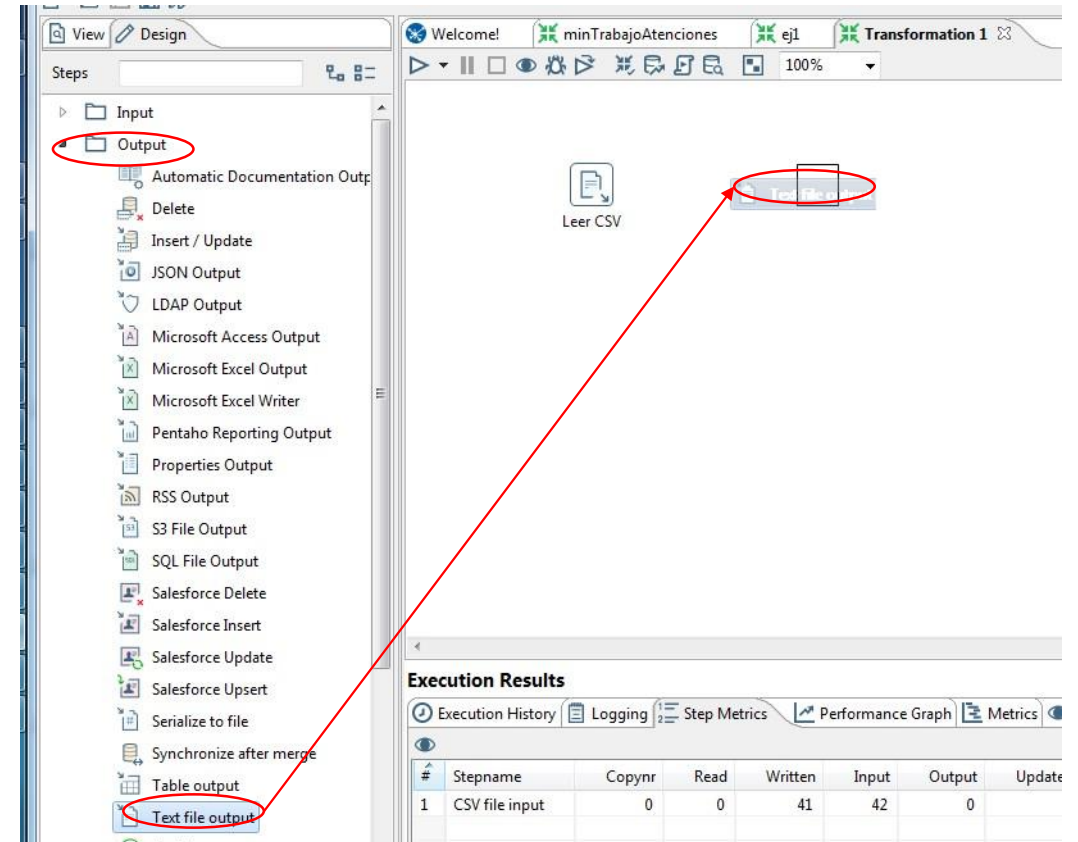
Help OK Get Fields Preview Cancel

CSV



# CARGA(L) CSV

- Muy similar a la extracción.



# FLUJO DE DATOS

- Conectar la extracción(Leer CSV) con la carga (Nuevo elemento).



Alternativamente: Mantener Shift, Hacer Drag and Drop desde origen a destino.

# UBICACIÓN DE CARGA

Text file output

Step name: Guardar CSV

File Content Fields

Filename: C:\Users\asierra\Google Drive\javeriana\Gestion de Datos\Ejemplo\codigosPostalesMedDestino Browse...

Run this as command instead? ☐

Pass output to servlet ☐

Create Parent folder ☒

Do not create file at start ☐

Accept file name from field? ☐

File name field:

Extension: csv

Include stepnr in filename? ☐

Include partition nr in filename? ☐

Include date in filename? ☐

Include time in filename? ☐

Specify Date time format ☐

Date time format:

Show filename(s)...

Add filenames to result ☒

Help OK Cancel



# OPCIONES

Text file output

Step name: Guardar CSV

File Content Fields

Append ☐

Separator: ; Insert TAB

Enclosure: "

Force the enclosure around fields? ☐

Disable the enclosure fix? ☐

Header ☒

Footer ☐

Format: CR+LF terminated (Windows, DOS)

Compression: None

Encoding: UTF-8

Right pad fields ☐

Fast data dump (no formatting) ☐

Split every ... rows: 0

Add Ending line of file

Help OK Cancel



# SELECCIONAR CAMPOS

Text file output

Step name: Guardar CSV

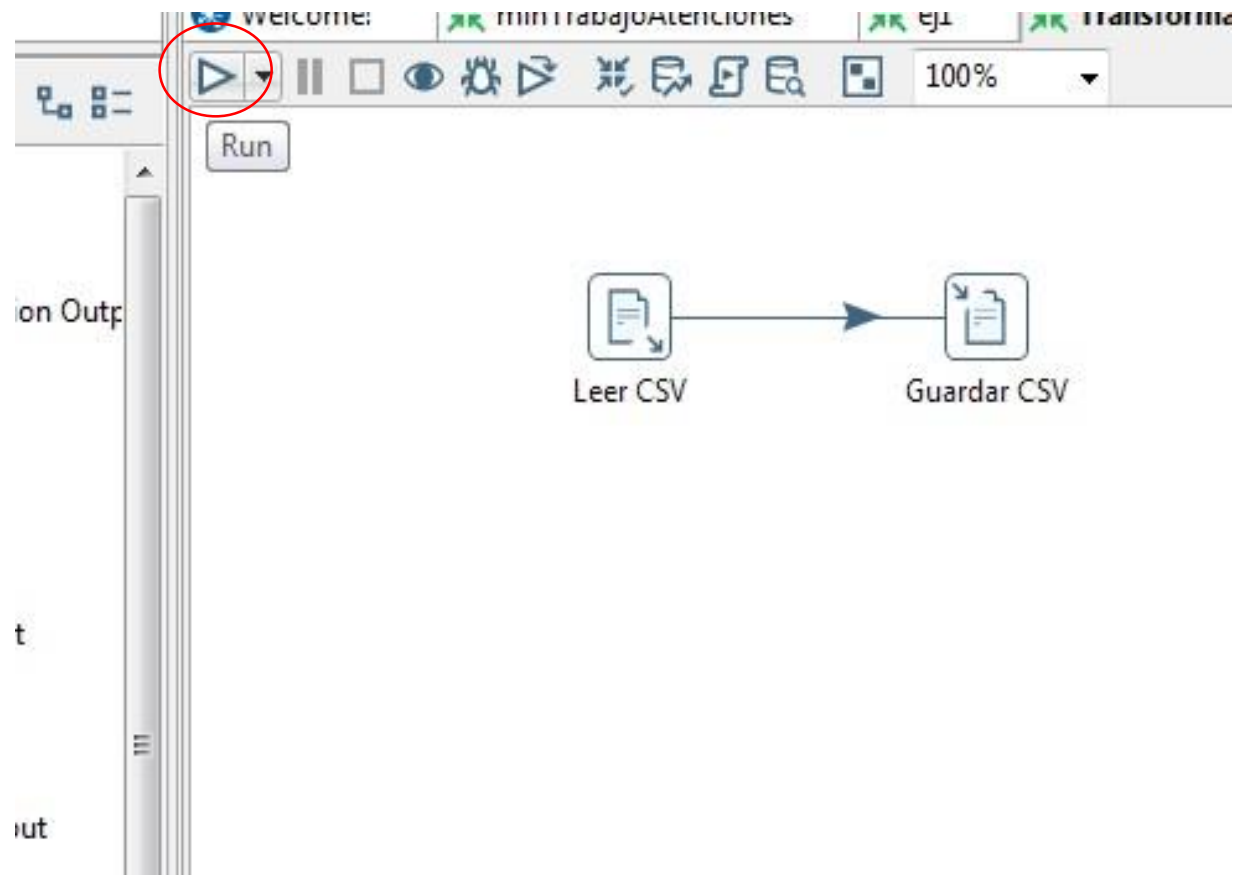
File Content Fields

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group
1	codigo_departamento	Integer	#	15	0	\$	,	.
2	nombre_departamento	String		9				
3	codigo_municipio	String	#	15				
4	nombre_municipio	String		8				
5	codigo_postal	String	#	15				
6	tipo	String		6				
7	barrios_contenidos_en_el_codigo_postal	String		254				

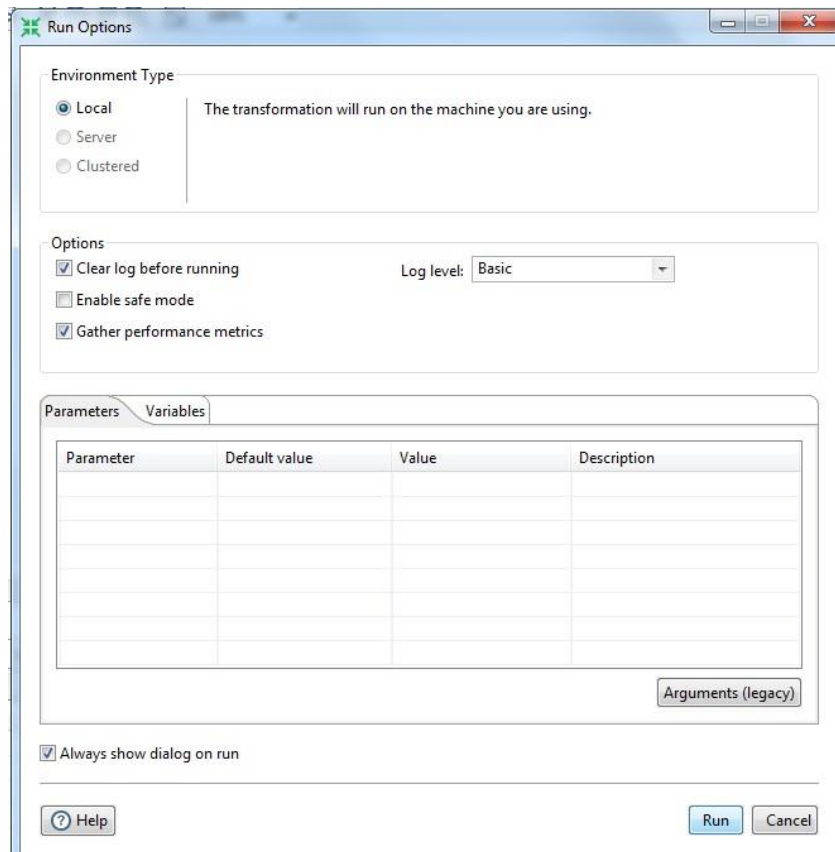
Get Fields Minimal width

Help OK Cancel

# EJECUTAR ETL

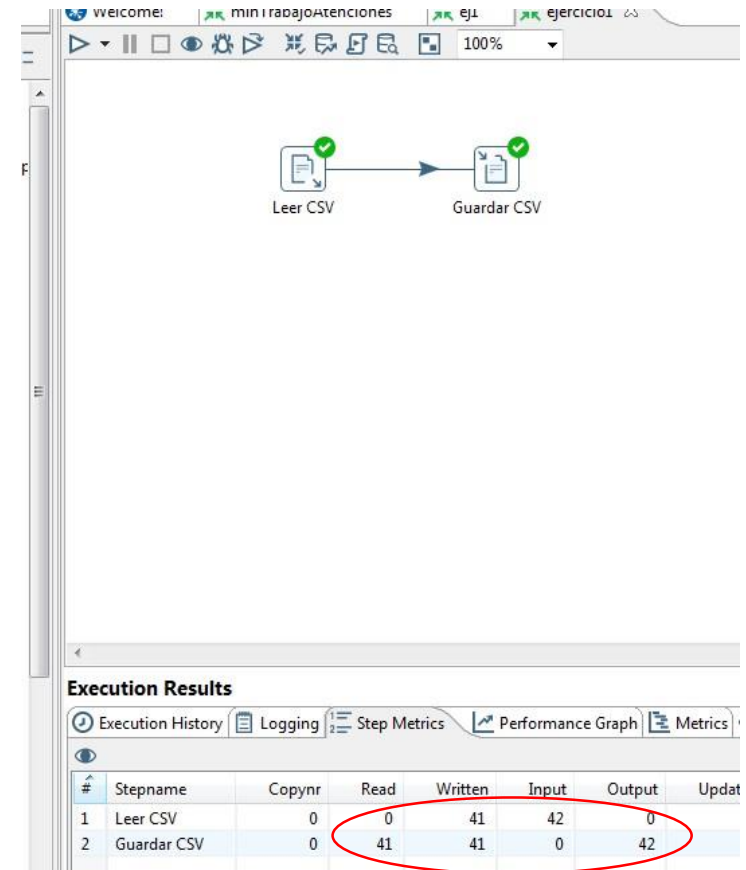


# EJECUCIÓN Y RESULTADO



The Run Options dialog box is shown with the following settings:

- Environment Type:** Local (selected). The transformation will run on the machine you are using.
- Options:**
  - ☒ Clear log before running
  - ☐ Enable safe mode
  - ☒ Gather performance metrics
- Log level:** Basic
- Parameters/Variables:** A table with 4 columns: Parameter, Default value, Value, and Description. It is currently empty.
- Arguments (legacy):** A button to view legacy arguments.
- ☒ Always show dialog on run
- Buttons:** Help, Run, and Cancel.



The Execution Results window shows the execution history of the transformation. The top section displays a flow diagram with two steps: 'Leer CSV' and 'Guardar CSV', both marked with green checkmarks. The bottom section shows the 'Execution Results' table.

#	Stepname	Copynr	Read	Written	Input	Output	Updat
1	Leer CSV	0	0	41	42	0	
2	Guardar CSV	0	41	41	0	42	

# CONEXIÓN A UNA BASE DE DATOS (DBMS)

- Cada proveedor tiene una estructura ligeramente diferente.
- Normalmente.
  - Dirección IP o Nombre del Servidor.
  - Usuario.
  - Contraseña.
  - Base de Datos.
  - Puerto.

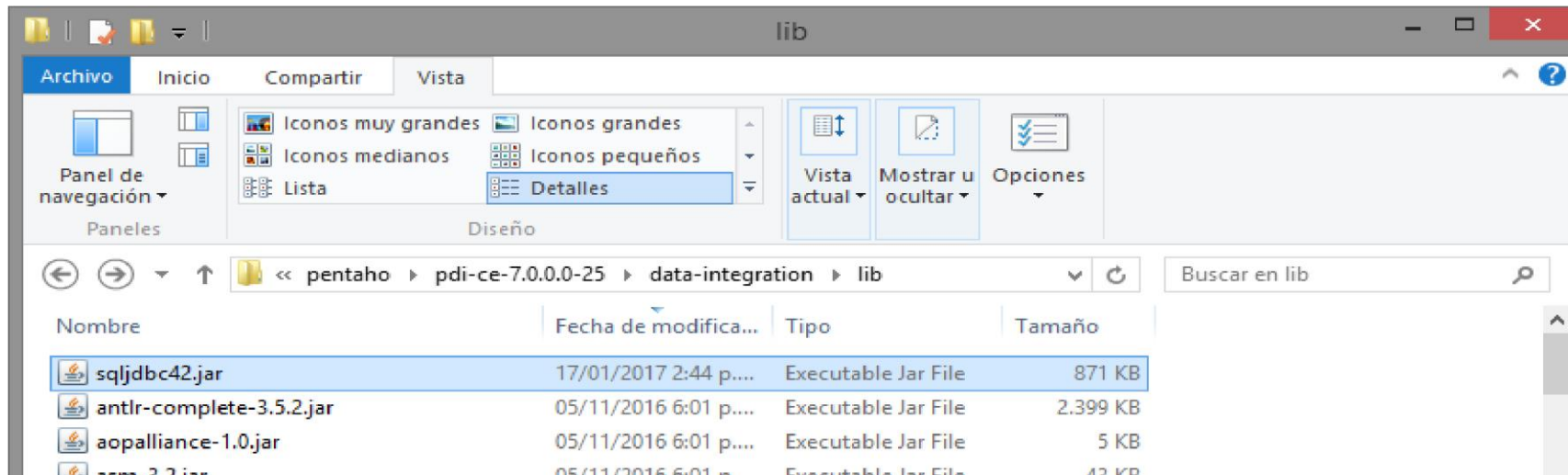
# SQL SERVER – SQL MANAGEMENT STUDIO

- Conectarse.
  - Estos son los datos para la clase.
  - Cada escenario es diferente.
  - Normalmente un DBA otorga estos datos.



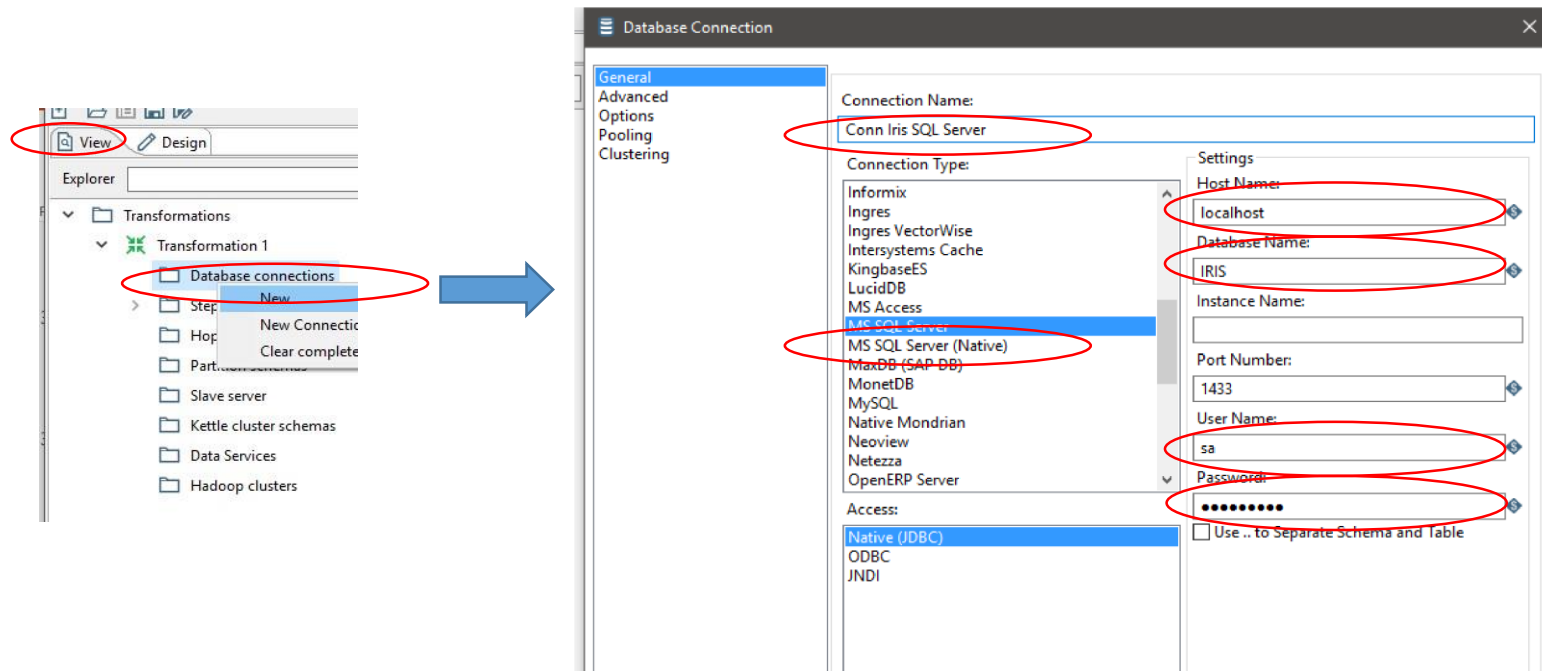
# INSTALACIÓN DE DRIVER SQL SERVER

- Pentaho no incluye driver para SQL Server.
- Se debe descargar de:
- <https://docs.microsoft.com/enus/sql/connect/jdbc/microsoft-jdbc-driver-for-sql-server>
- El archivo sqljdbc42.jar se copia en la carpeta lib de Pentaho.
  - Por ejemplo: C:\software\pdi-ce-7.0.0.0-25\data-integration\lib

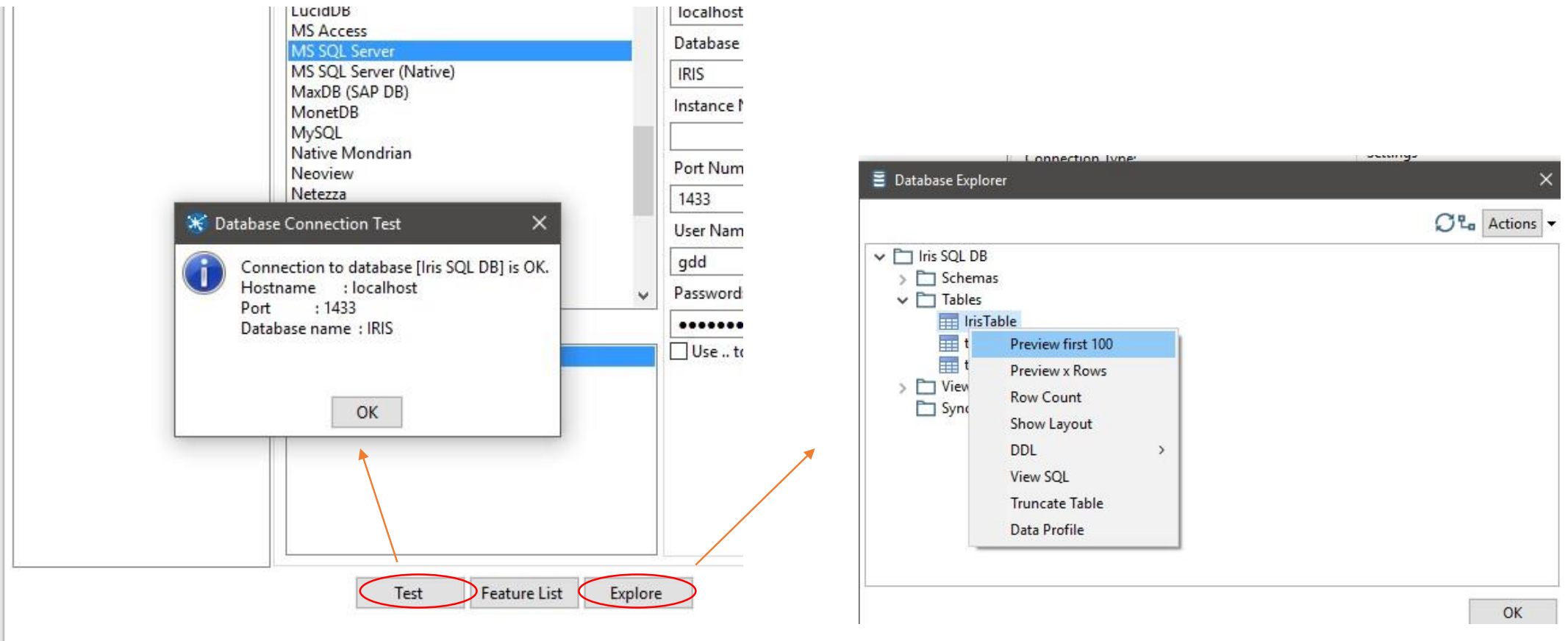


# CREAR CONEXIÓN A DBMS (SQL SERVER)

- Previo a interactuar con la Base de Datos, se debe crear la conexión desde Pentaho.



# PROBAR Y EXPLORAR





# EJERCICIO

- Extracción: Tabla de SQL Server (Cualquiera de AdventureWorks).
- Carga: csv.
- Estandar para archivos csv.
  - Separador:”,”
  - Con cabecera.
  - Con comillas dobles para texto.

# ENLACES DE INTERÉS - REFERENCIAS

- Documentación de Pentaho Data Integration (Kettle).
  - <http://wiki.pentaho.com/display/EAI/Latest+Pentaho+Data+Integration+%28aka+Kettle%29+Documentation>
- Datasets públicos.
  - <https://github.com/caesar0301/awesome-public-datasets>
  - <https://www.data.gov/>
- Integración de Datos.
  - Data integration blueprint and modeling techniques for a scalable and sustainable architecture, Anthony Giordano, IBM Press Pearson, 2011
    - Disponible en la biblioteca
- The Art of Enterprise Information Architecture
  - A Systems-Based Approach for Unlocking Business Insight
  - Mario Godinez, Eberhard Hechler, Klaus Koenig, Steve Lockwood, Martin Oberhofer, Michael Schroeck
  - IBM Press
  - 2010