Componente ETL

José Samos Jiménez

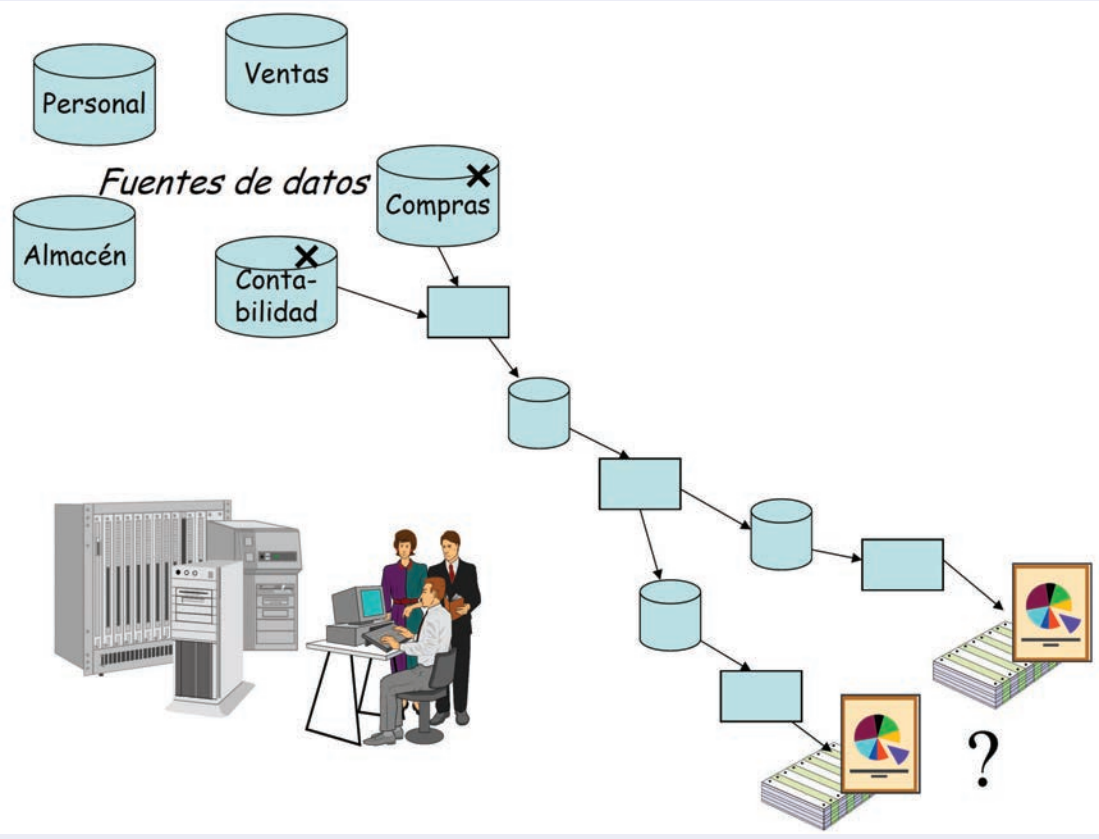
2020 jsamos (lsi-ugr)
Departamento de Lenguajes y Sistemas Informáticos
Universidad de Granada

Curso 2019-20

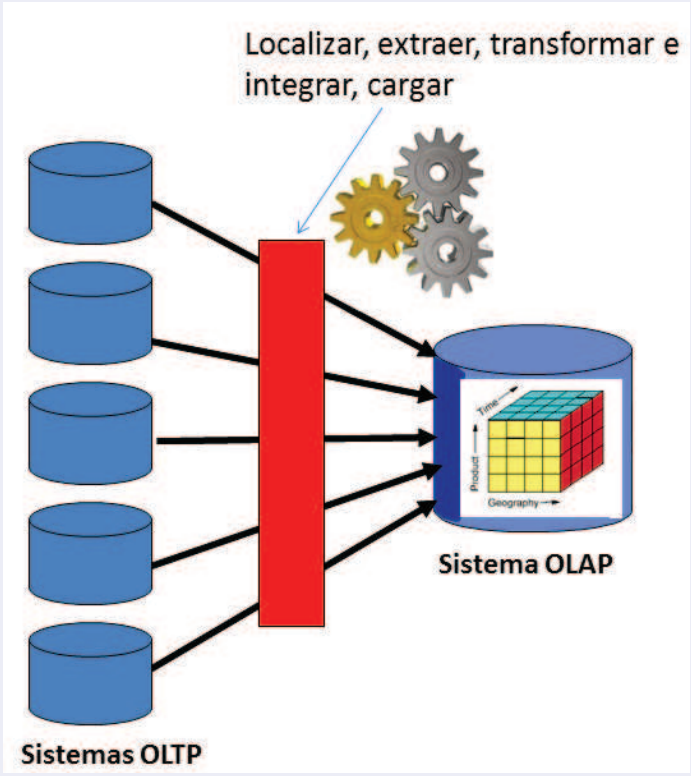
Contenido

Componente ETL y modelo lógico

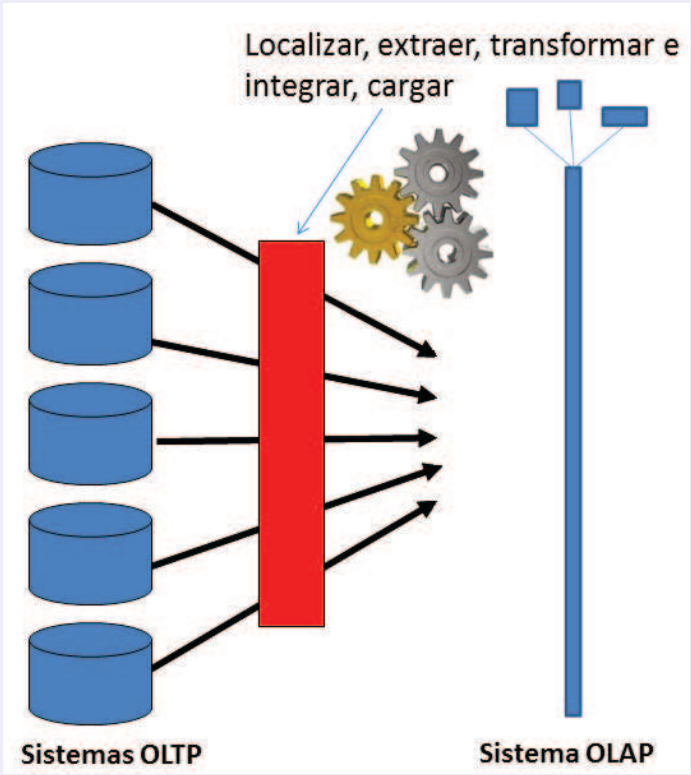
ETL para un informe



Componente ETL

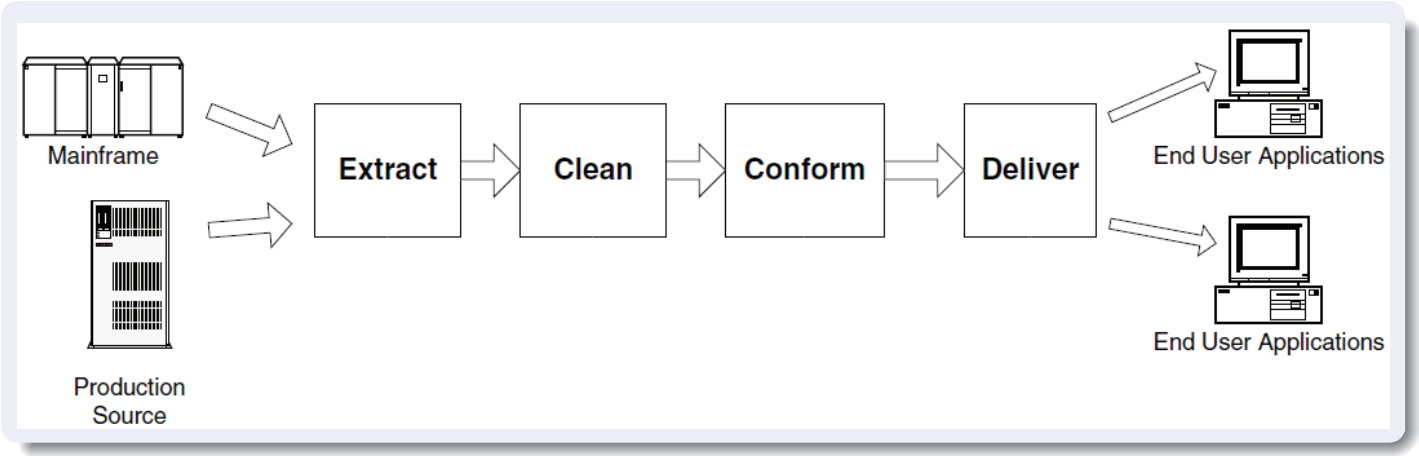


Componente ETL y modelo lógico



Proceso de replicación de datos

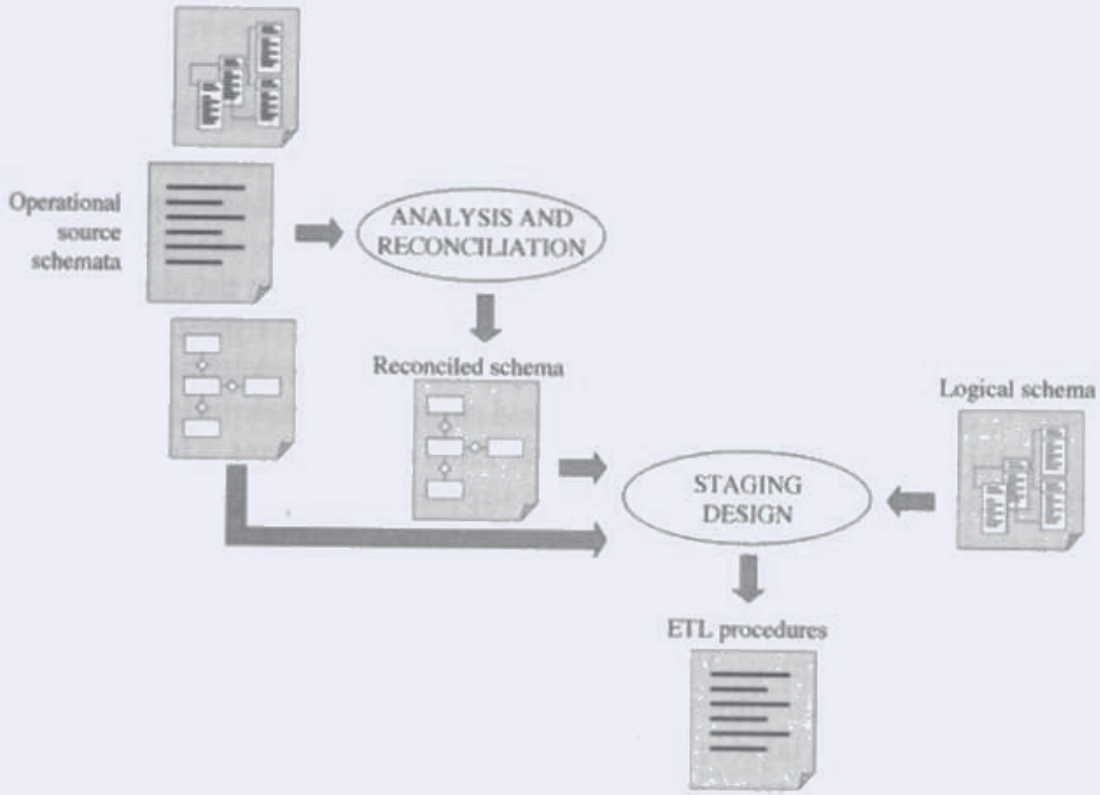
Flujo de datos



Proceso de replicación

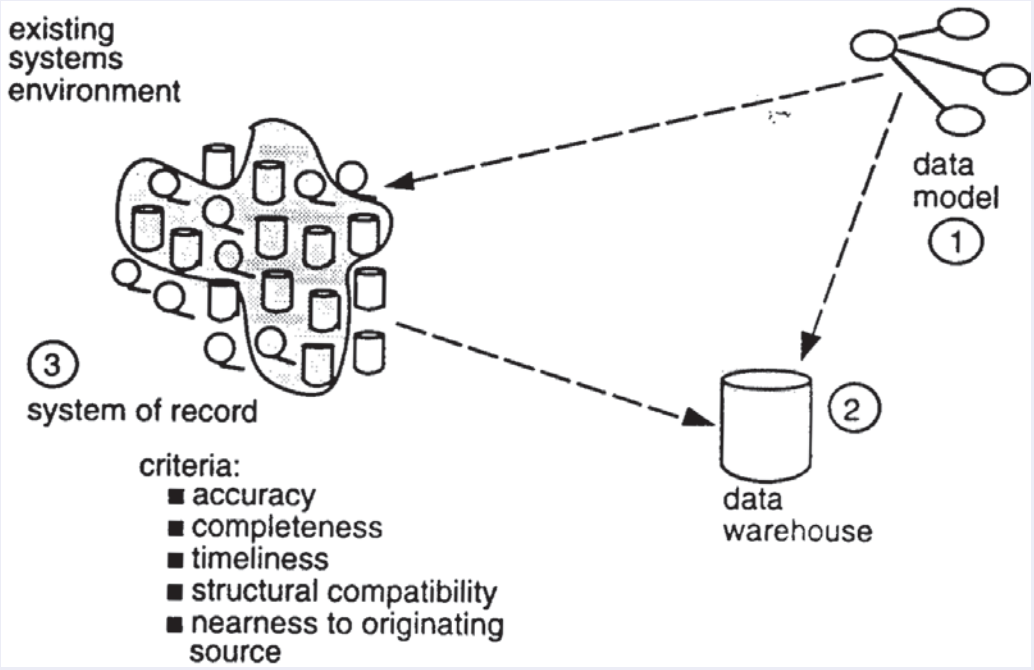
Function	Process step	Frequency
Administration	1. Identify the source data.	Once
Administration	2. Identify or define the target data.	Once
Administration	3. Create the mapping between source and target.	Once
Administration	4. Define the replication mode.	Once
Administration	5. Schedule the process of replication.	Once
Capture	6. Capture the required data from the source.	Frequently
Data transfer	7. Transfer the captured data between source and target.	Frequently
Transformation	8. Transform the captured data based on the defined mapping.	Frequently
Apply	9. Apply the captured data to the target.	Frequently
Process management	10. Confirm the success or failure of the replication.	Frequently
Process management	11. Document the outcome of the replication in the metadata.	Frequently
Administration	12. Maintain the definitions of source, target, and mapping.	As needed

Fuentes de datos y destino



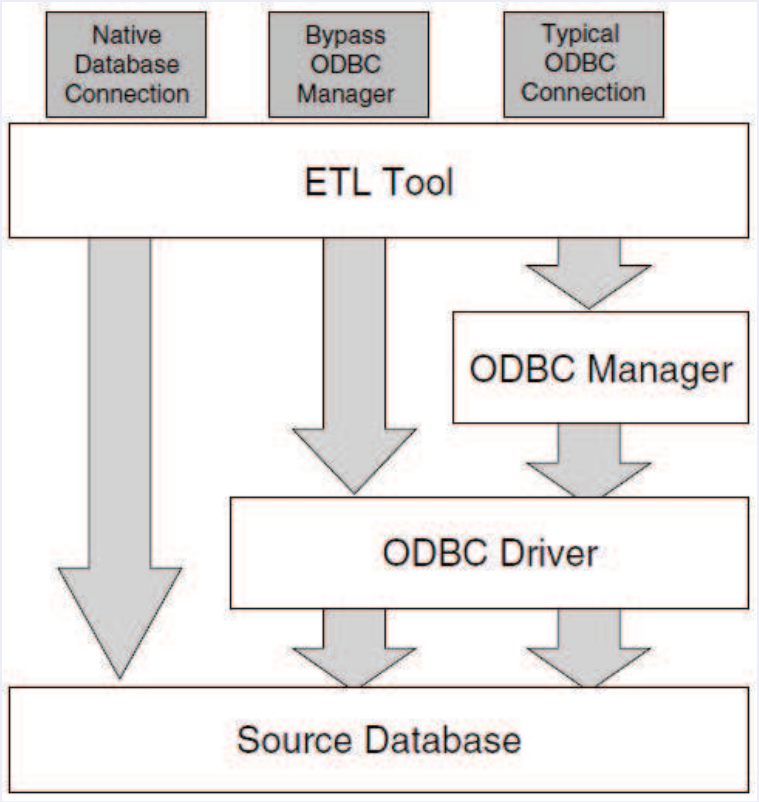
Fuentes de datos

Definir el *sistema fuente de registro*



- *System of record* (*SoR*) o *source system of record* (*SSoR*)

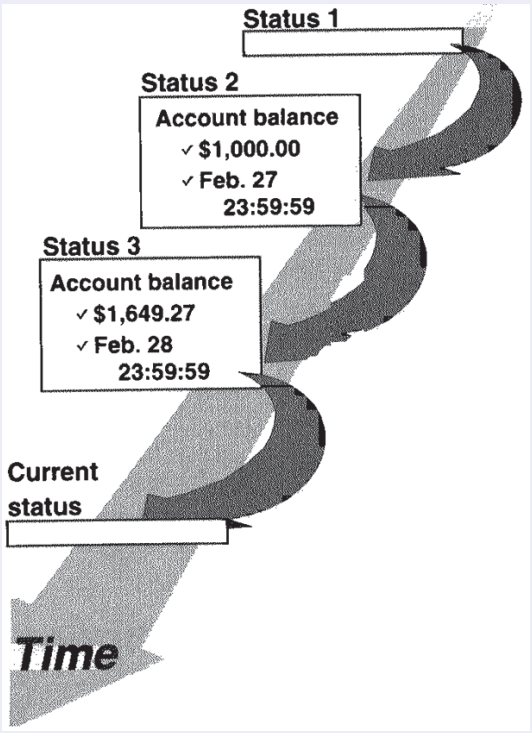
Acceso a las fuentes



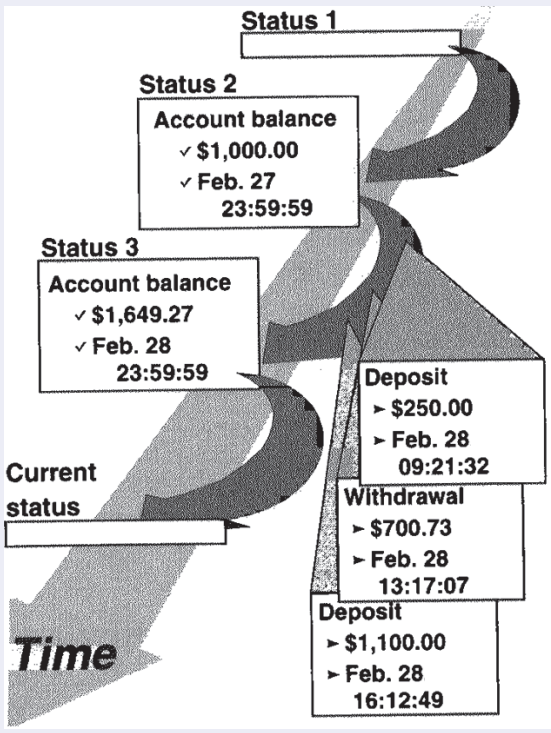
Extracción

Tipos de métodos

Diferidos

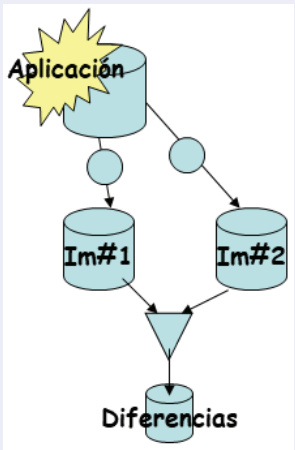


Inmediatos

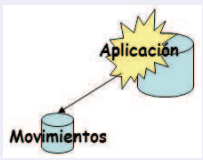


Métodos de extracción

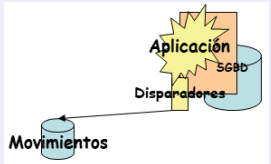
Comparación de imágenes



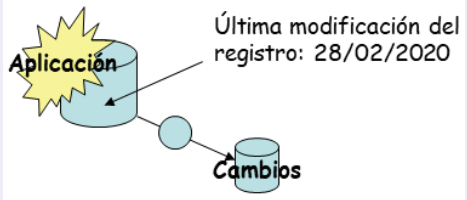
Generados por las aplicaciones



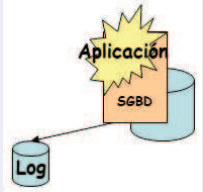
Mediante disparadores



Huella de tiempo



Log file



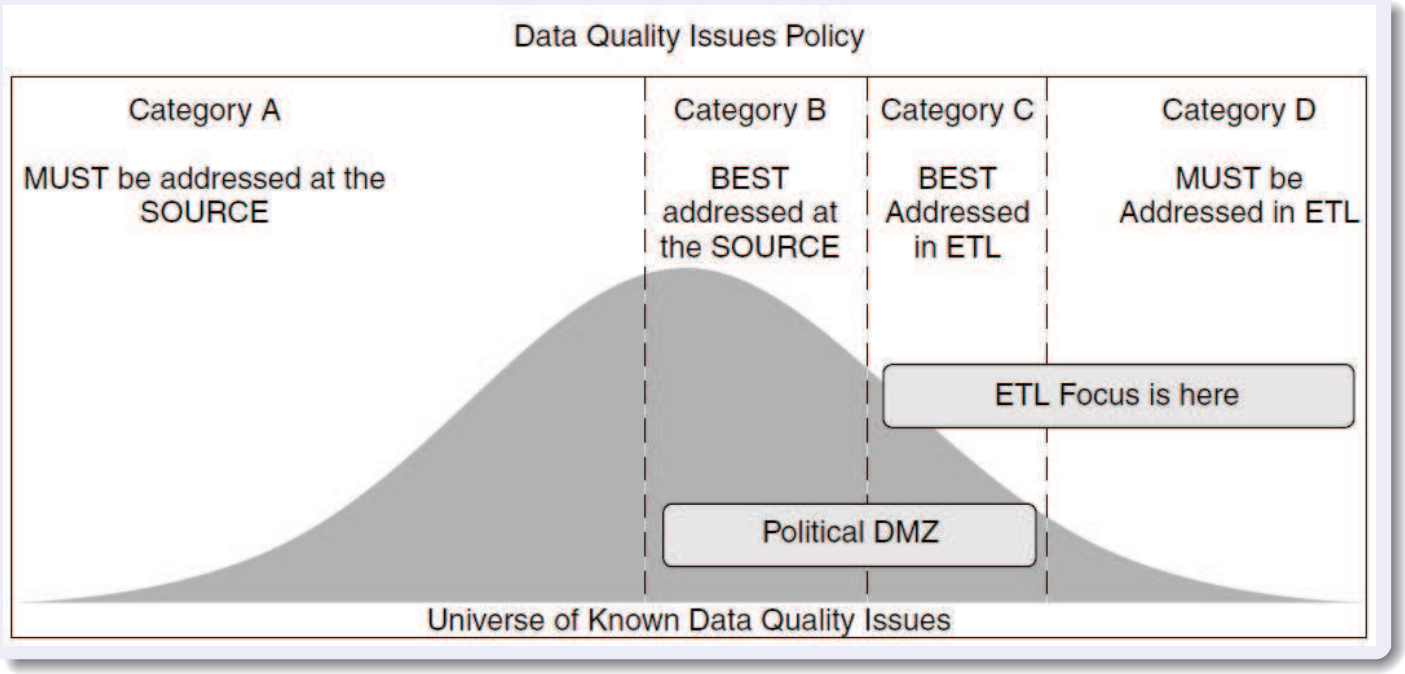
Transformación

Transformación de los datos

- Fusionar datos de varias fuentes.
- Adaptar los datos al modelo destino:
  - Cambiar la codificación, el formato, la granularidad... de los datos.
- Problemas: datos repetidos, inconsistentes o erróneos, datos que faltan...

- Notificar los problemas y errores a las fuentes.

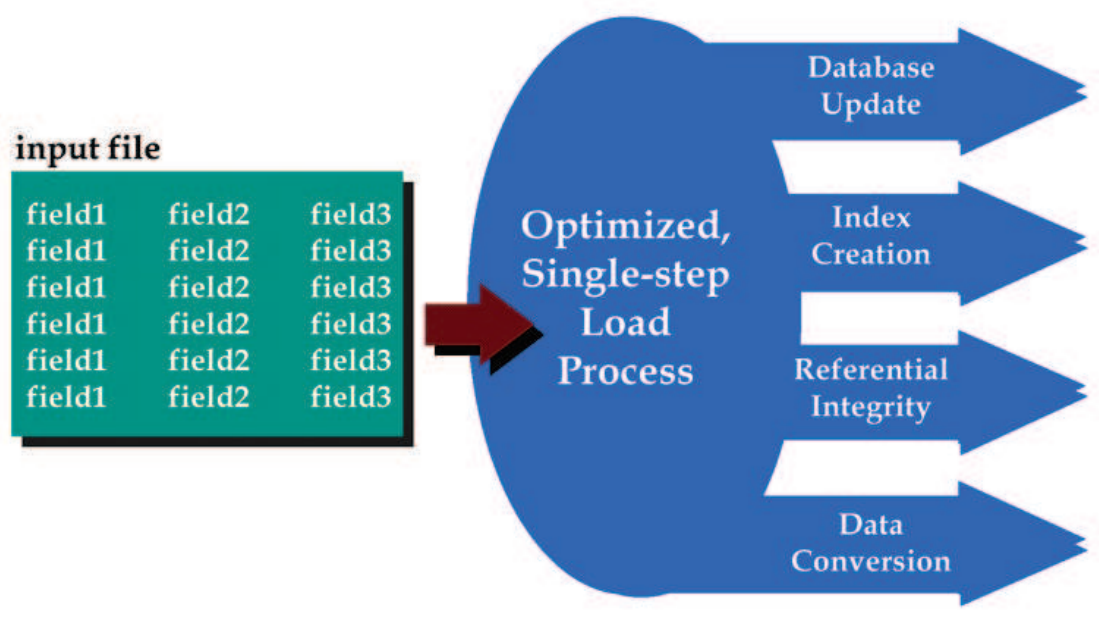
Corrección de problemas y errores



- Frontera: *Political DMZ* (zona desmilitarizada).

Carga

Cargar los datos

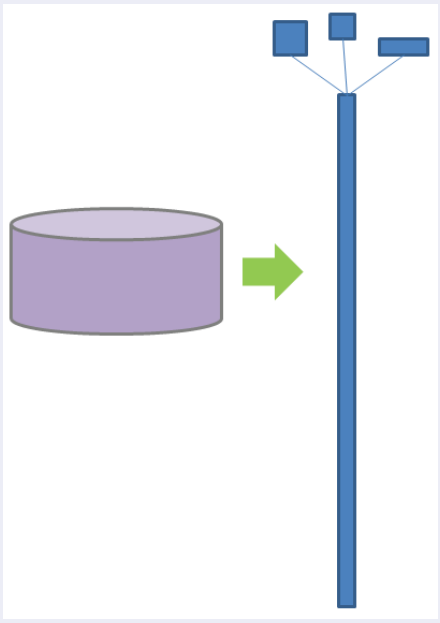


- Sumarizaciones.
- (*SCD*: *dimensiones lentamente cambiantes*.)

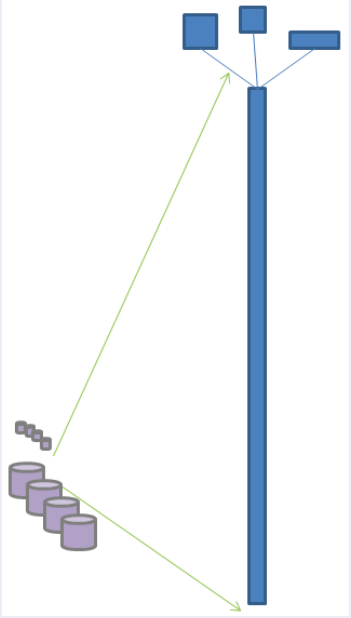
Periodicidad de las actualizaciones

Carga inicial y actualizaciones periódicas

Carga inicial



Actualizaciones periódicas

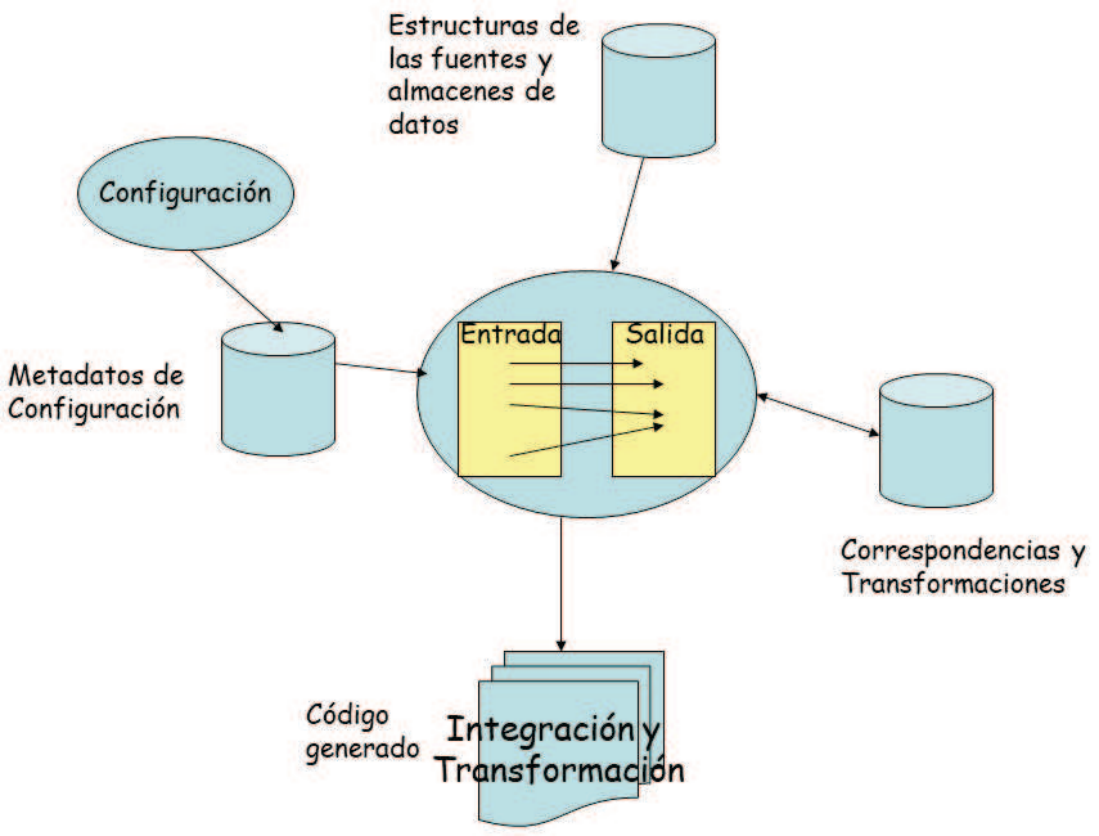


- Definir la periodicidad del proceso de replicación.

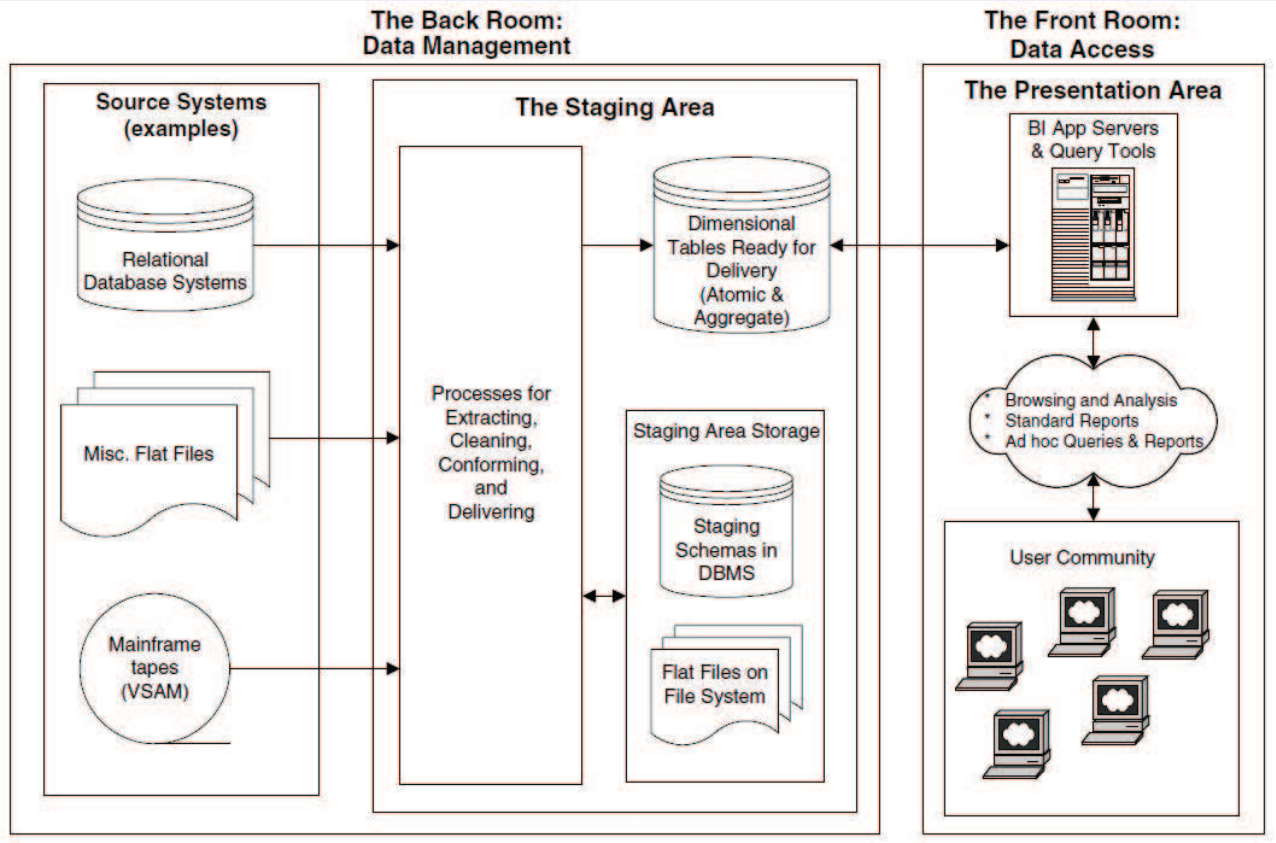
Herramientas ETL y arquitectura

Estructura general

Herramientas ETL

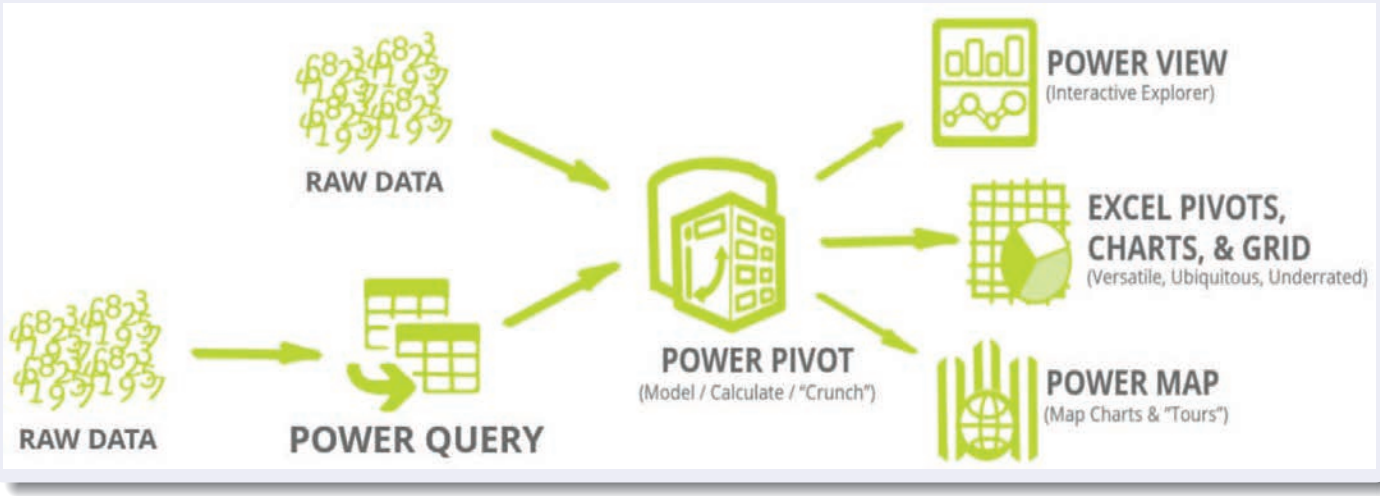


«Mercado - cocina - comedor» [KC04]



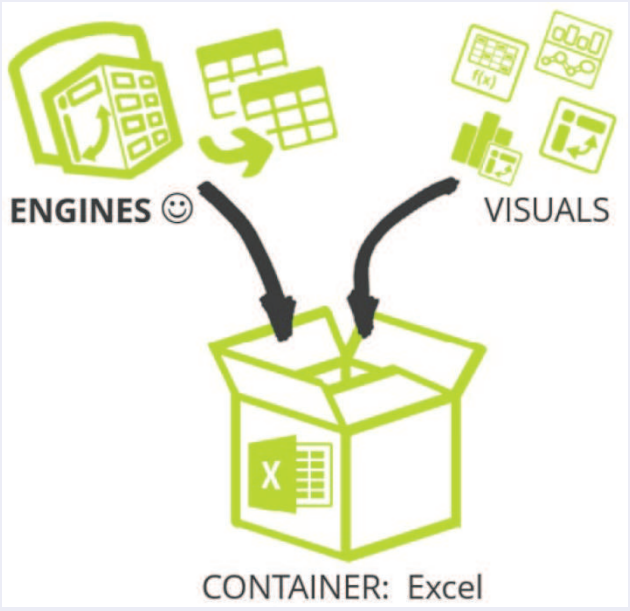
ETL de usuario final: *Power Query*

Autoservicio de Transformación y de Sistemas OLAP

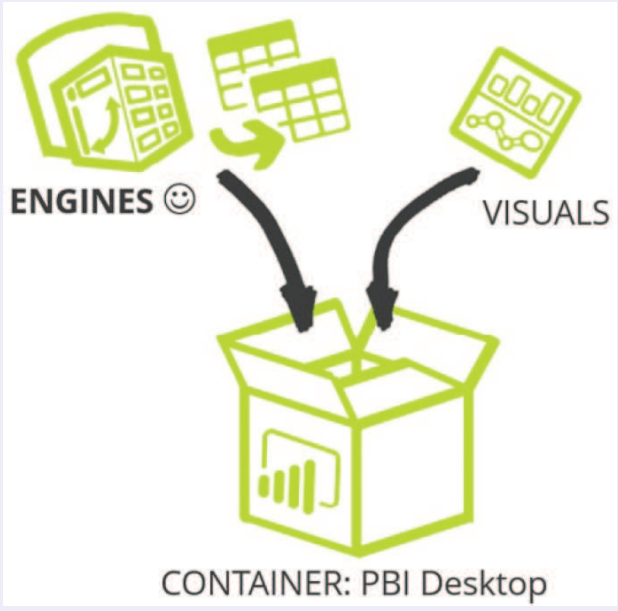


Power Query y *Power Pivot* en *Excel* y en *Power BI*

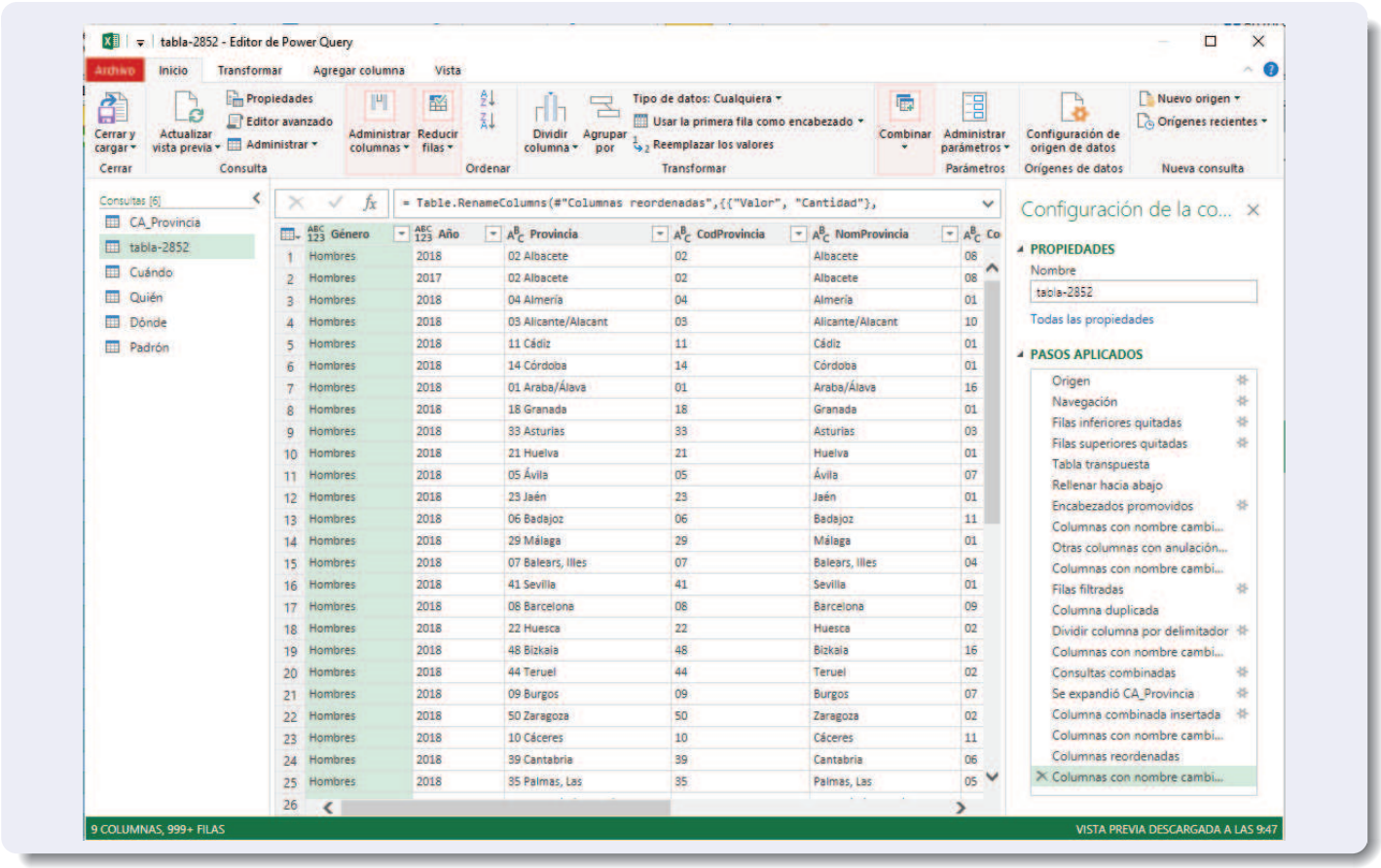
Excel



Power BI

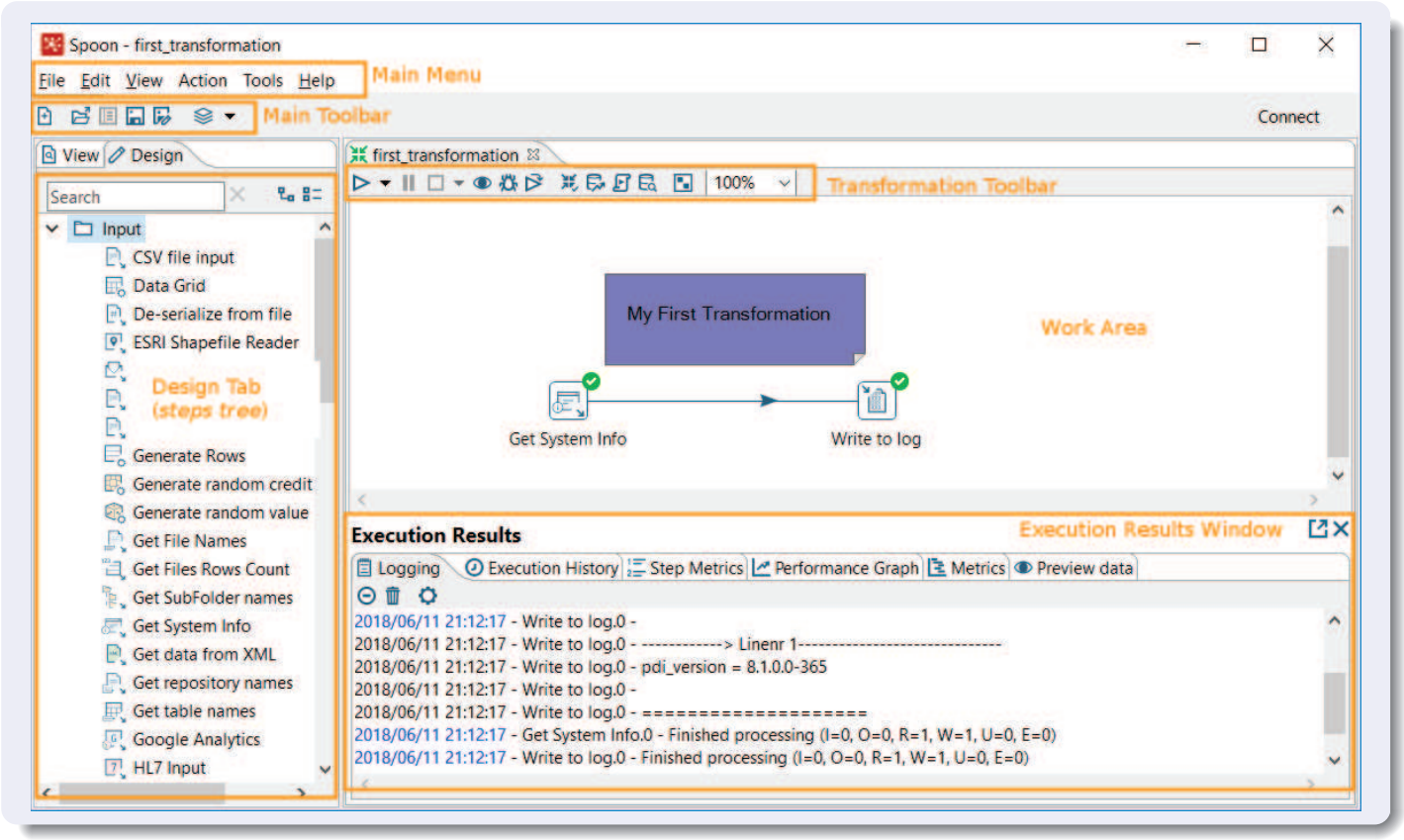


Transformación de datos (*Power Query*)

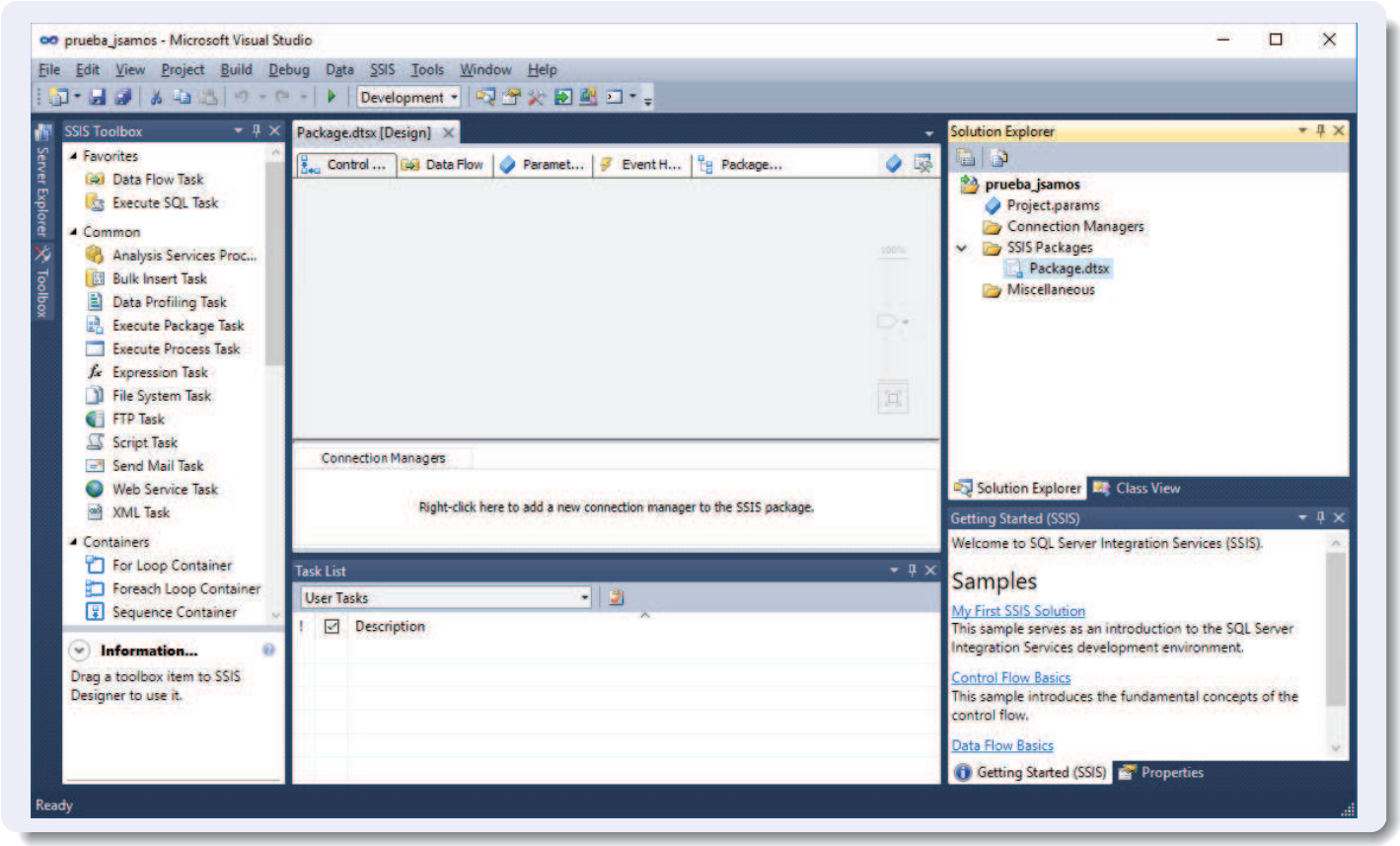


ETL profesional

PDI (Pentaho Data Integration): Spoon



SSIS (SQL Server Integration Services)



Bibliografía

Bibliografía

GR09 M. Golfarelli, S. Rizzi. *Data Warehouse Design: Modern Principles and Methodologies*. McGraw-Hill, 2009.

JPT10 C. Jensen, T. Pedersen, C. Thomsen: *Multidimensional Databases and Data Warehousing*. Morgan & Claypool, 2010.

KC04 R. Kimball, J. Caserta: *The Data Warehouse ETL Toolkit*. Wiley, 2004.