

# ETL

Master Data Analytics para para la empresa

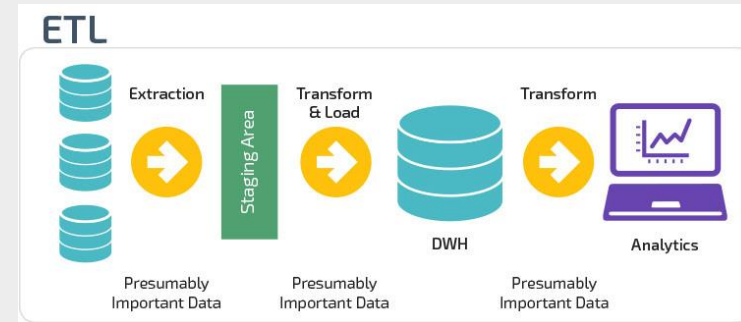
Pedro Nieto

# Agenda

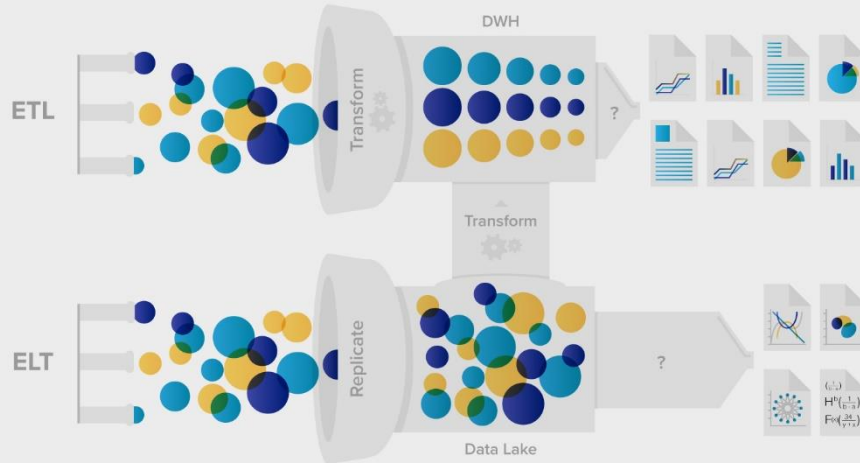
- 1. Qué es ETL?**
- 2. ETL vs ELT**
- 3. Zonas de trabajo**
- 4. ETL (Fases)**
- 5. Game Players**

## ¿Qué es una ETL?

Es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

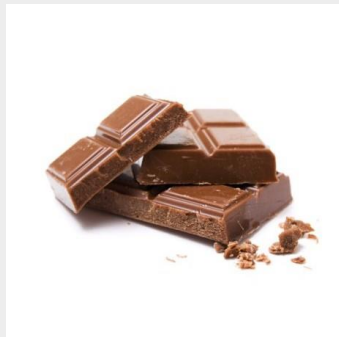
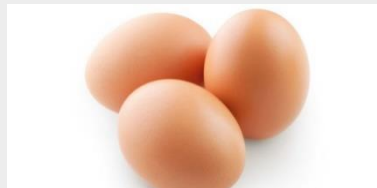


# ETL vs ELT



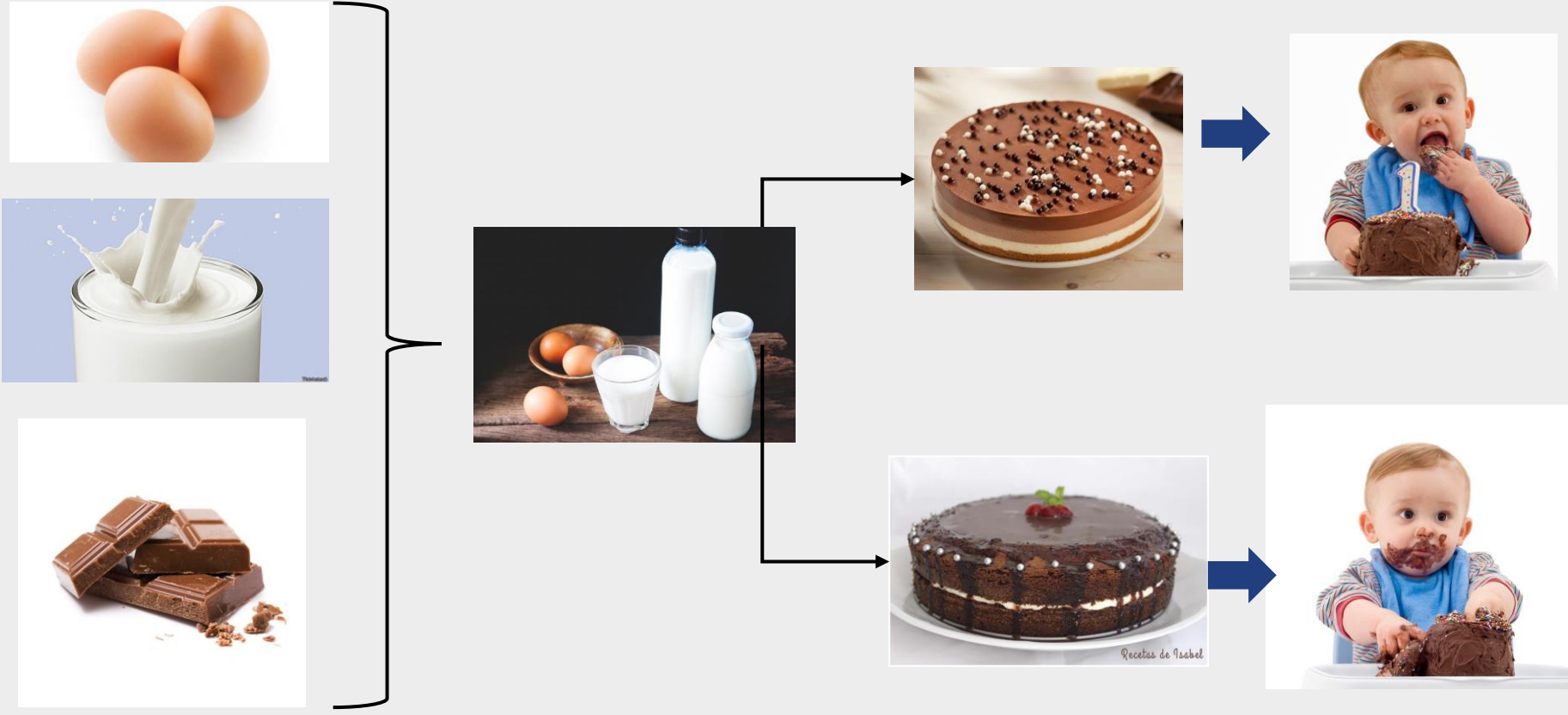
- **With ETL** you extract data, transform it into a compatible structure, and load it into a target data warehouse system—so business intelligence tools can query and analyze it.
- **With ELT** you extract data, immediately load it into the target data lake system, and then transform the data—so business intelligence tools can query and analyze it. ELT generally refers to the data transformation process required to transform data already located in a data lake or warehouse.

## Más sencillo... Hagamos una tarta con una ETL

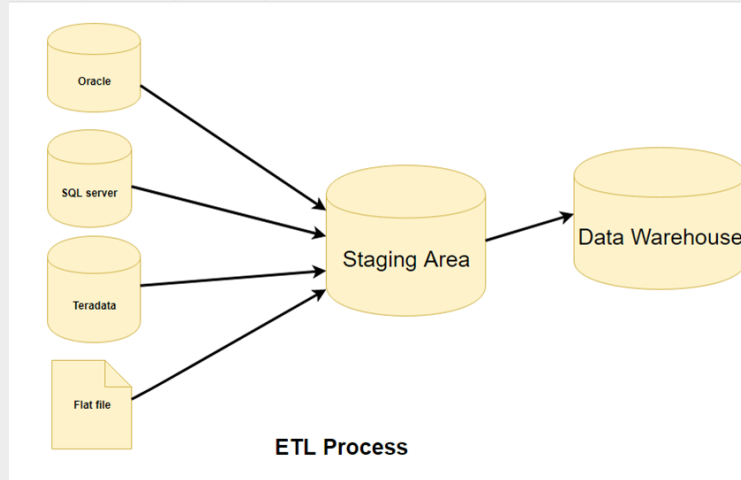




# Más sencillo... Hagamos una tarta con una ELT



# Zonas de Trabajo



La existencia de una zona de Staging es una practica necesaria en el trabajo con ETLs.

Esta zona proporciona una serie de beneficios:

- Aislamiento de origen
- Consolidación de fuentes
- Alineamiento con Reference Data
- Limpieza de Datos
- Trazabilidad completa

## ETL (Extracción)

En este paso, los datos se extraen del sistema de origen en el área de staging. Las transformaciones, si las hay, se realizan en el área de staging para que el rendimiento del sistema de origen no se degrade.

### Modos de extracción

Básicamente, existen tres modos distintos de extracción. El tipo de necesidad de la organización es lo que, normalmente, determinará la elección de una u otra forma.

#### *Full Extract o extracción total*

Esta modalidad consiste en extraer la totalidad de datos. En este caso, se barren tablas completas que pueden llegar a tener millones de registros.

#### *Incremental Extract o extracción incremental*

Se va procesando por lotes únicamente lo que fue modificado o agregado. También puede haber filas que se borren por estar duplicadas, tratarse de datos erróneos, etc.

#### *Update Notification o notificación de actualizaciones*

En este caso, solo se van extrayendo los datos a medida que se produce una actualización (por ejemplo, un inserto) .

Estos tres tipos de extracción son manejados por un módulo denominado *Change Data Capture* (CDC).



# Connectors

Over 100 connectors



# ETL (Transformación)

La fase de transformación de un proceso de ETL aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados. Estas directrices pueden ser declarativas, pueden basarse en excepciones o restricciones pero, para potenciar su pragmatismo y eficacia, hay que asegurarse de que sean:

- Declarativas.
- Independientes.
- Claras.
- Inteligibles.
- Con una finalidad útil para el negocio.



## ETL (Load)

La fase de carga es el momento en el cual los datos de la fase anterior (**transformación**) son cargados en el sistema de destino.

•**Acumulación simple:** La acumulación simple es la más sencilla y común, y consiste en realizar un resumen de todas las transacciones comprendidas en el período de tiempo seleccionado y transportar el resultado como una única transacción hacia el data warehouse, almacenando un valor calculado que consistirá típicamente en un sumatorio o un promedio de la magnitud considerada.

•**Rolling:** El proceso de **Rolling** por su parte, se aplica en los casos en que se opta por mantener varios niveles de granularidad. Para ello se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones de la unidad de tiempo o diferentes niveles jerárquicos en alguna o varias de las dimensiones de la magnitud almacenada (por ejemplo, totales diarios, totales semanales, totales mensuales, etc.).



# Pros and Cons

## ETL

<b>Time-Load</b>	High
<b>Time-Transformation</b>	High once – Low multiple
<b>Time- Maintenance</b>	High maintenance if data changes
<b>Implementation Complexity</b>	Learning curve is easier

## ELT

Low
Low once – High Multiple
Low maintenance as data is always available.
Complex at origin



# Market Players

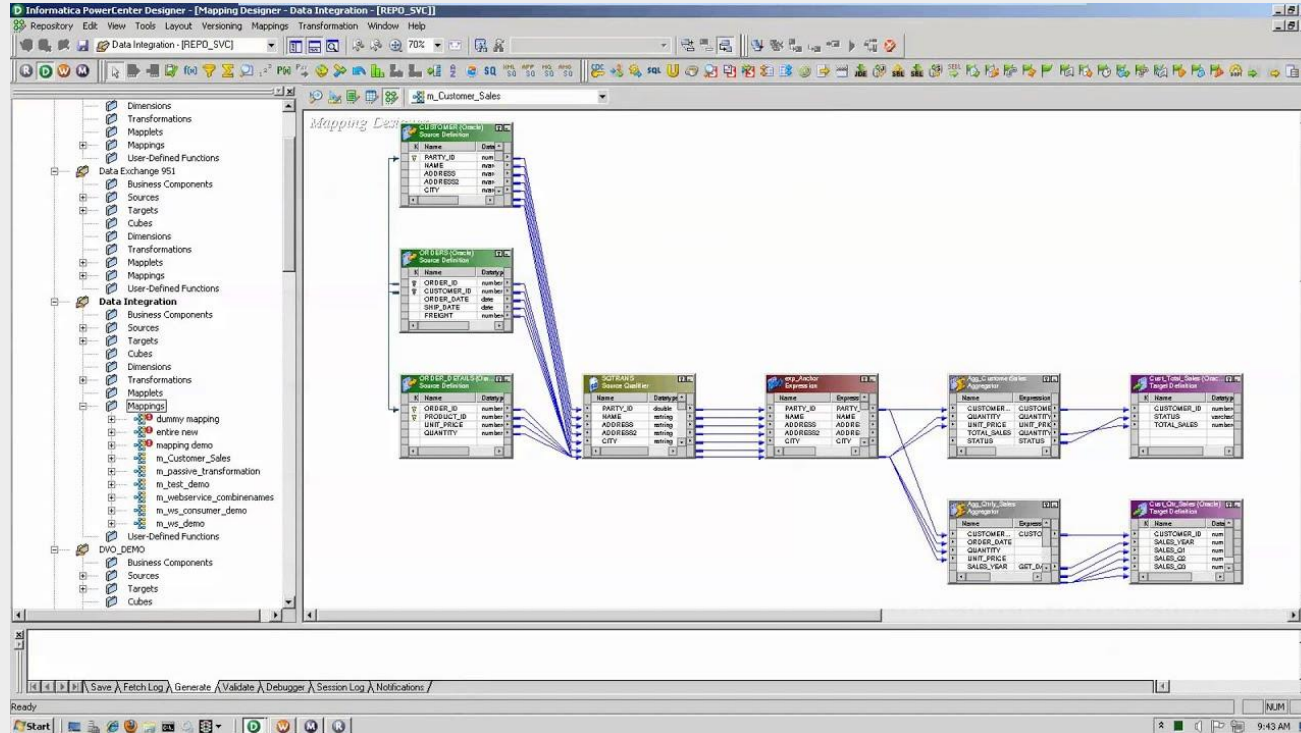
Figure 1. Magic Quadrant for Data Integration Tools



Source: Gartner (August 2019)

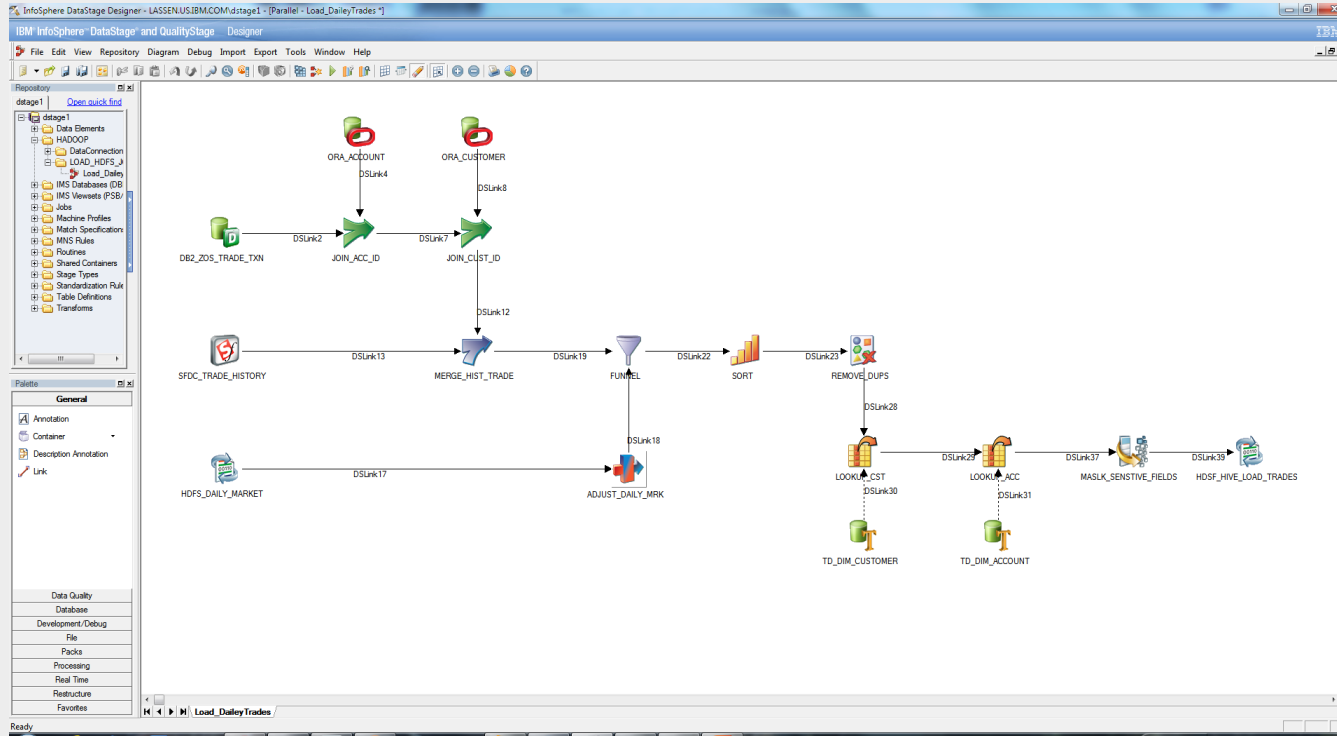


# Informatica Powercenter





# IBM Data Stage



# Oracle Data Integrator

The screenshot displays the Oracle Data Integrator 11g interface. The main window shows a data flow diagram with two source nodes connected to a target node. The left pane shows the project structure, including a folder named 'JOINTEST\_MODEL\_FOLDER' containing a 'JOIN\_SOURCE' model. The right pane shows the 'Target Datastore - TRG\_REGION\_JOINTEST' mapping table.

**Target Datastore - TRG\_REGION\_JOINTEST**

Position	Indicators	Name	Mapping
1		*REGION_ID	SRCREG.REGION_ID
2		COUNTRY_ID	SRCCNT.COUNTRY_ID
3		REGION	SRCREG.REGION
4		COUNTRY	SRCCNT.COUNTRY

**Diagram - Property Inspector**

Grid Size: 15  
Show Grid: False

# Talend Data Integration

The screenshot displays the Talend Open Studio for Data Integration (5.6.2.20150508\_1414) interface. The main workspace shows a data integration job named 'updatecredit 0.1'. The job flow is as follows:

- Repository:** The left pane shows the 'LOCAL: cloud' repository with various components like 'loan\_investor\_loan\_c 0.1', 'loan\_investor\_loan\_account\_txns\_c 0', 'loan\_loan\_account\_c 0.1', 'loan\_loan\_payment\_transaction\_c 0', 'loan\_other\_transaction\_c 0.1', 'loan\_repayment\_schedule\_c 0.1', 'loan\_repayment\_schedule\_summary', 'sample 0.1', 'sfdc 0.1', 'updated 0.1', and 'updatecredit 0.1'.
- Job Flow:**
  - The job starts with a **tSalesforceConnection\_1** component.
  - It then branches into two parallel paths:
    - Path 1:** **tSalesforceGetServerTimestamp\_1** (row4 (Main)) → **tJavaRow\_2**.
    - Path 2:** **tInfiniteLoop\_1** (Iterate) → **tJava\_1** → **tSalesforceGetUpdated\_1** (row2 (Main)).
  - The paths merge at a **tMap\_1** component.
  - The output of **tMap\_1** is split into two paths:
    - Path 3:** **tMysqlInput\_1** (row3 (Lookup)) → **tMysqlOutput\_1** (updates (Main order:1)).
    - Path 4:** **tMysqlOutput\_2** (row1 (Main)) → **tMysqlOutput\_2** (inserts (Main order:2)).
- Designer/Code:** The bottom pane shows the 'Designer' tab for the **tSalesforceGetUpdated\_1** component. The 'Basic settings' section is visible, showing 'Use an existing connection' checked, 'Authentication' set to 'Basic', and 'Login Type' set to 'Basic'.



## Ejercicio1:

- Debéis leer un fichero CSV y escribirlo a fichero Json en la misma carpeta:
- Hacer lo mismo con MASTERS.csv
- Leer de un json y escribir en un CSV

### Entregables:

1) Captura de pantalla de código y fichero

## Ejercicio2:

- Debéis leer un fichero CSV(alumnos) y reemplazar:
  - China -> CH
  - France -> FR
  - Brazil -> BR
  - Cuba -> CU
- Debéis leer un fichero CSV(alumnos) y filtrar, de forma que solo me devuelva los alumnos que son de Portugal.
- Leer CSV ALUMNOS, reemplazar Brazil por BR, y filtrar solo los de BR

### Entregables:

1) Captura de pantalla de código y fichero



## Ejercicio3:

- Leer tabla de actores y volcarlo a fichero JSON
- [Talend Metadata DB - YouTube](#)
  - Para ello hay que conectarse a la base de datos dvdrental

DB Type --> PostgreSQL

DB Version --> v9 and later

String of connection: jdbc:postgresql://35.192.88.224:5432/dvdrental?

Login --> postgres

Password --> Welcome01

Server --> 35.192.88.224

Port -> 5432

Database --> dvdrental

Schema --> public

### Entregables:

**1) Captura de pantalla de código y fichero**

## Ejercicio4:

- Agregar las películas por rating y mostrar un count, volcar a json el resultado

### Entregables:

1) Captura de pantalla de  
código y fichero

## Ejercicio5:

- Realizar un Join entre Actor / Film / Film\_Actor y volcar a json un fichero con estos campos:
  - Nombre
  - Apellido
  - Película

### Entregables:

1) Captura de pantalla de código y fichero

## Ejercicio6:

- Cargar en un csv la cantidad de dinero Gastada por usuario, nombre y apellido:

### Entregables:

- 1) Captura de pantalla de código y fichero

## Ejercicio7:

- Cargar en un csv el numero de veces que se ha alquilado cada pelicula

### Entregables:

- 1) Captura de pantalla de código y fichero

## Ejercicio 8:

- Import from BigQuery

[Ejercicio 8 Import data from BigQuery with Talend - YouTube](#)

**Entregables:**  
**1) Captura de  
pantalla de código  
y fichero**