EDEM Escuela de Empresarios

GFT

ETL

Master Data Analytics para para la empresa

Pedro Nieto

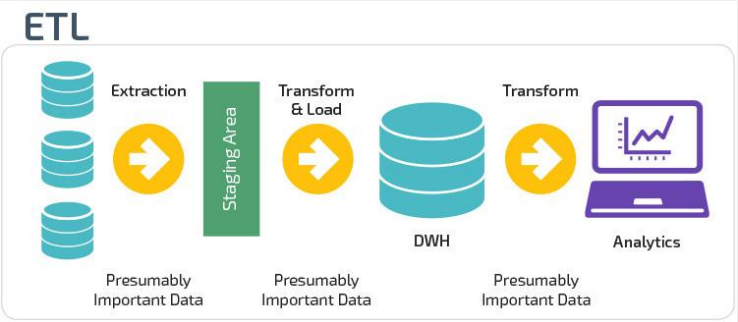
EDEM Escuela de Empresarios

Agenda

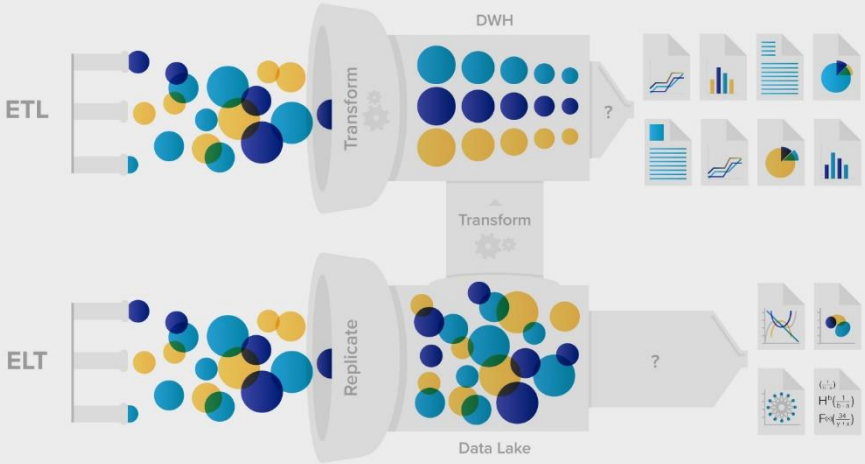
1. Qué es ETL?
2. ETL vs ELT
3. Zonas de trabajo
4. ETL (Fases)
5. Game Players

¿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



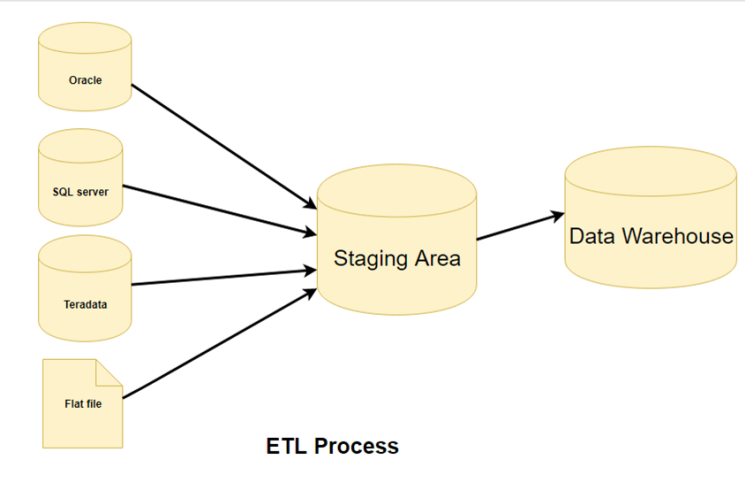
- **With ETL** you extract data, transform it into a compatible structure, and load it into a target data warehouse system—so business intelligence tools can query and analyze it.

- **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

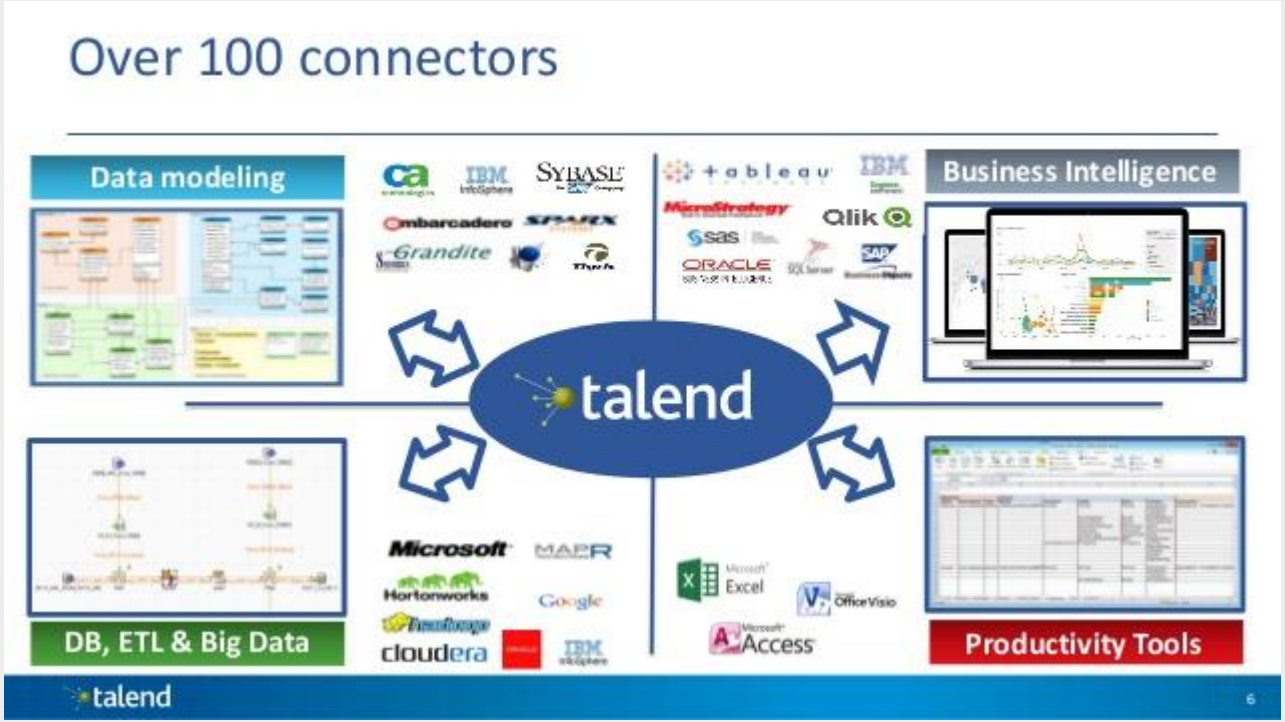
Se va procesando por lotes únicamente lo que fue modificado o agregado. También puede haber filas que se borren por estar duplicadas, tratarse de datos erróneos, etc.

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
Data modeling
Business Intelligence
talend
DB, ETL & Big Data
Productivity Tools
talend
6

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.).

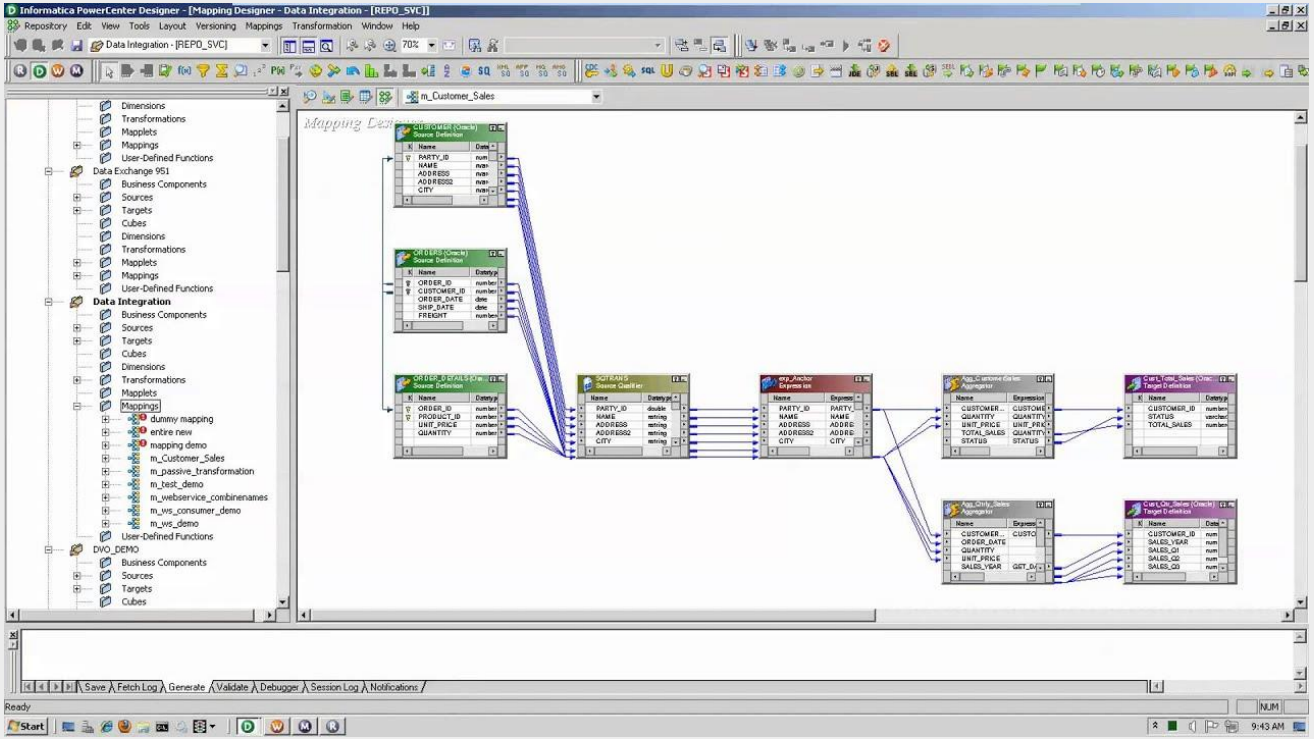
Pros and Cons

	ETL	ELT
Time-Load	High	Low
Time-Transformation	High once – Low multiple	Low once – High Multiple
Time- Maintenance	High maintenance if data changes	Low maintenance as data is always available.
Implementation Complexity	Learning curve is easier	Complex at origin

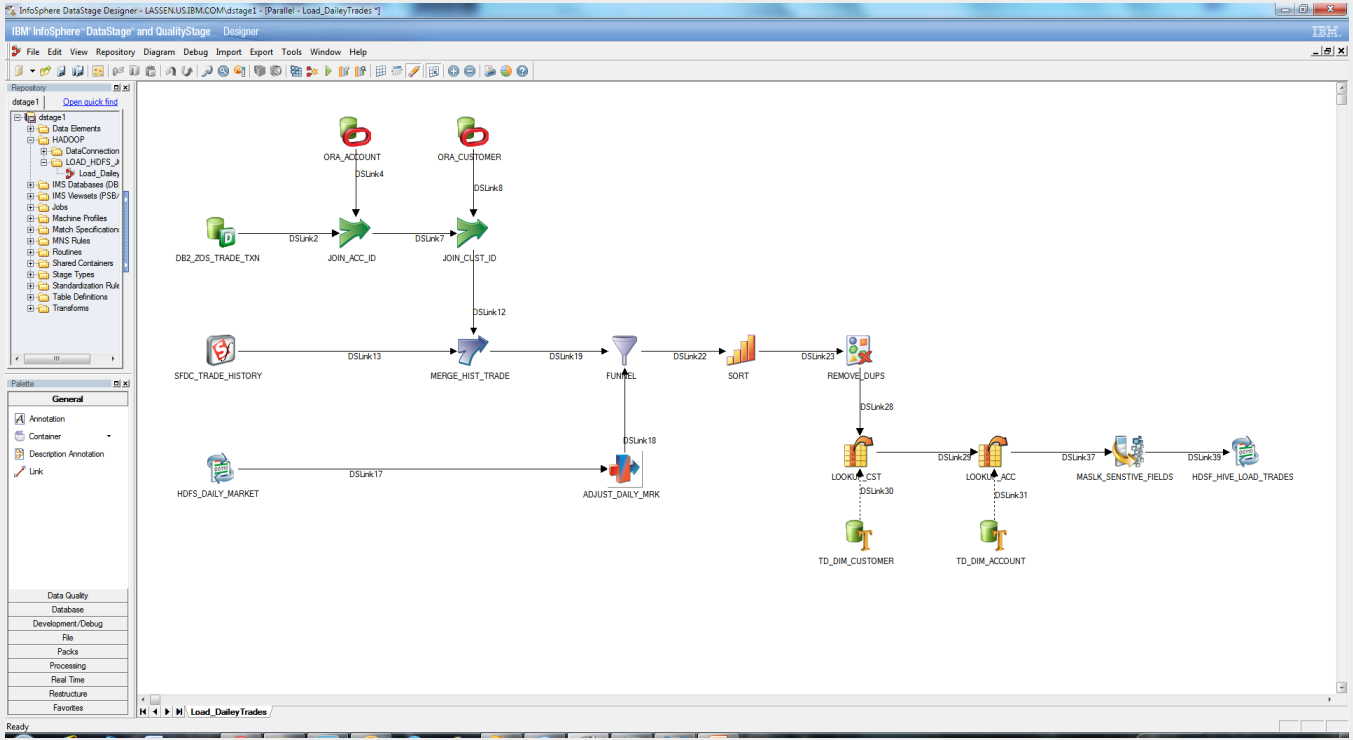
Market Players

Source: Gartner (August 2019)

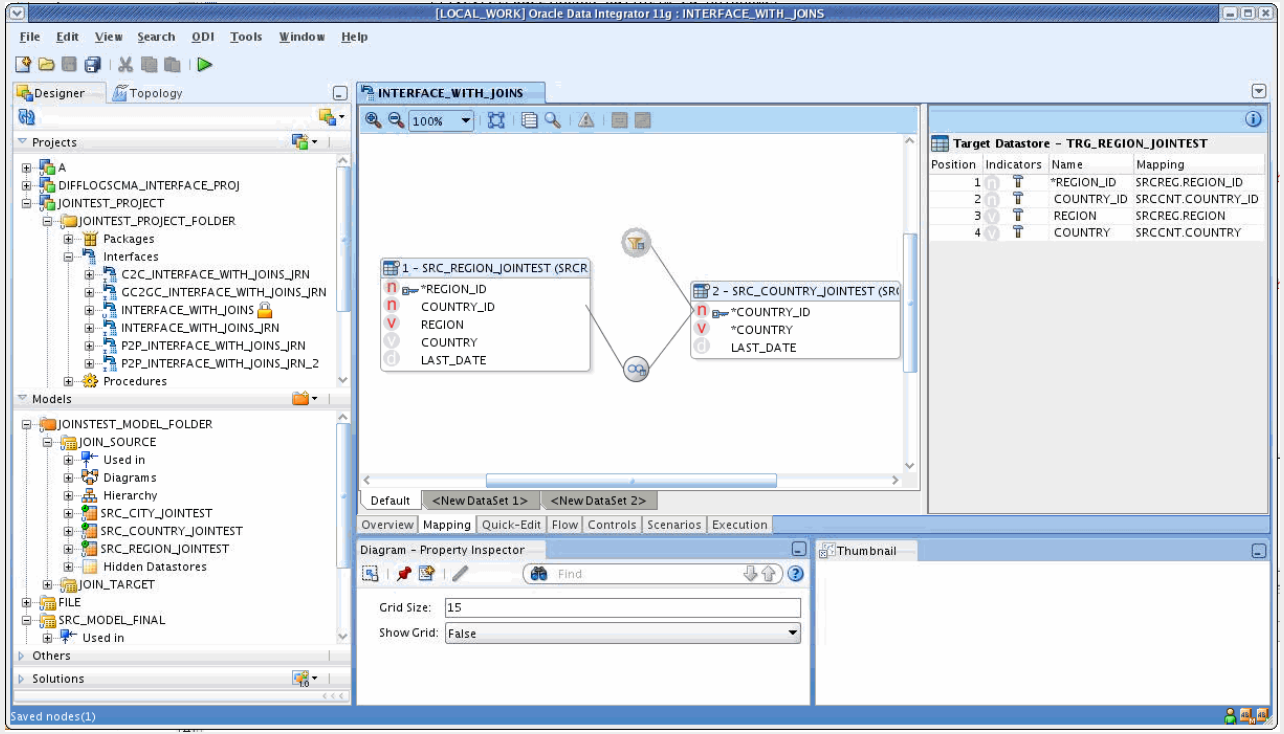
Informatica Powercenter



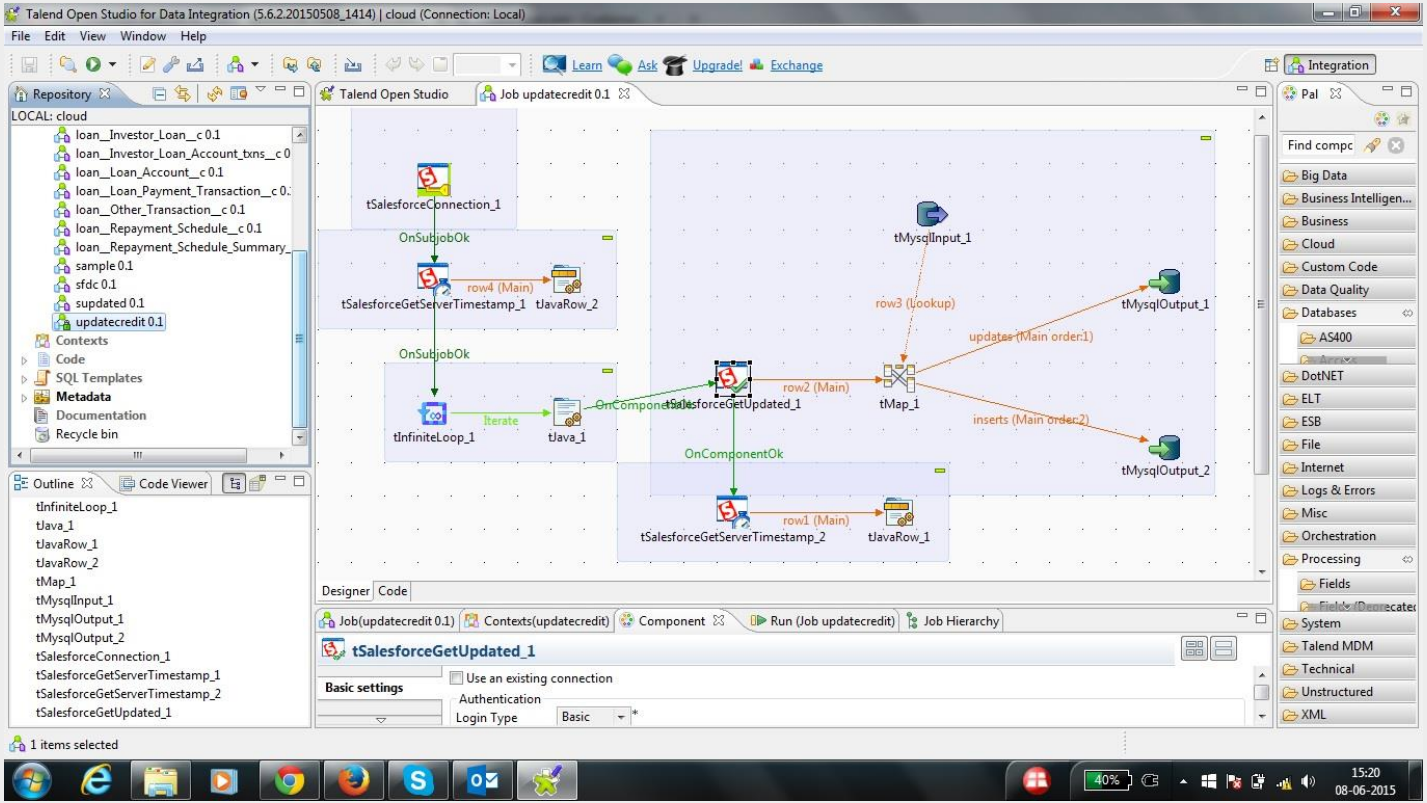
IBM Data Stage



Oracle Data Integrator



Talend Data Integration




LET'S
CODE

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