

# Caso práctico: almacén de datos para el análisis de los registros de viajes de la New York City Taxi and Limousine Commission (NY taxis)

## PR2 – Carga de datos

**Autores:**        **David Díaz Arias**  
                      **Carles Llorach i Rius**  
                      **Maria Nerea Sevilla Marchena**

## Índice

<b>1. Presentación</b>	<b>2</b>
1.1. Descripción	2
1.2. Criterios de evaluación	2
1.3. Guía de muestra	3
<b>2. Identificación de los procesos ETL</b>	<b>3</b>
<b>3. Diseño y desarrollo de los procesos ETL (60 %)</b>	<b>5</b>
3.1. Creación de tablas	5
3.2. Establecer el entorno	6
3.3. Conexiones	6
3.4. Bloque IN	8
<b>4. Implementación de los trabajos con procesos ETL (20%)</b>	<b>12</b>
<b>5. Formato y fecha de entrega</b>	<b>13</b>

# 1. Presentación

La PR2 consiste en la carga y transformación de datos relacionados con los registros de viajes de la TLC en un almacén de datos.

A partir de la solución oficial de la primera práctica (PR1), el estudiantado, de forma individual, debe diseñar, implementar y ejecutar los procesos de extracción, transformación y carga de los datos de las fuentes de datos proporcionadas.

Así pues, esta actividad tiene como objetivo identificar y desarrollar los procesos de carga del almacén de datos y que esta sea efectiva.

Para realizar la actividad, se cuenta con una serie de recursos de aprendizaje disponibles en el aula, junto al enunciado de la actividad. Se recomienda consultar el documento «Guía de buenas prácticas ETL», así como los vídeos proporcionados.

## 1.1. Descripción

Si nos centramos en los subobjetivos, esta segunda parte del caso práctico consiste en lo siguiente:

- 1) Identificar los procesos de extracción, transformación y carga de datos (ETL) hacia el almacén de datos.
- 2) Diseñar y desarrollar los procesos ETL mediante las herramientas de diseño proporcionadas.
- 3) Implementar con los trabajos (*jobs*) los procesos ETL para que su carga planificada sea efectiva.

Además del documento con la solución de la PR2 entregado, se tendrá en cuenta la implementación realizada sobre la máquina virtual proporcionada en el curso.

En concreto, el documento de la solución de la PR2 debe incluir los siguientes aspectos:

- Descripción de todas las acciones que se han realizado.
- Capturas de pantalla que demuestren todas las partes significativas del ETL, sus características y su correspondiente explicación.
- Capturas de pantalla que demuestren la correcta carga de los datos (cargados en la base de datos).

## 1.2. Criterios de evaluación

La nota final se calculará a partir de la suma de los siguientes apartados:

- Identificación de los procesos de ETL (20 %):
  - Identificación y descripción de los procesos ETL de «ORIGEN» a «STAGE» («BLOQUE IN») (10 %).
  - Identificación y descripción de los procesos ETL de «STAGE» a «DW» («BLOQUE TR\_DIM») (5 %).

- Identificación y descripción de los procesos ETL de «STAGE» a «DW» (BLOQUE TR\_FACT) (5 %).
- Diseño y desarrollo de los procesos ETL (60 %):
  - *Script* de creación de las tablas intermedias (*staging area*) (0 %).
  - *Script* de creación del modelo dimensional (0 %).
  - Uso de variable de entorno y configuración conexiones (5 %).
  - Transformaciones «Bloque IN» (cinco procesos) (15 %).
  - Transformaciones «Bloque TR\_DIM» (seis procesos) (20 %).
  - Transformaciones «Bloque TR\_FACT» (dos procesos) (20 %).
- Implementación de los trabajos (JOBS) de los procesos ETL (20 %):
  - *Jobs* «Bloque IN» (5 %).
  - *Jobs* «Bloque TR\_DIM» (5 %).
  - *Jobs* «Bloque TR\_FACT» (5 %).
  - Proceso completo (utiliza los *jobs* IN y TR) «Bloque DW» (5 %).

### 1.3. Guía de muestra

Con el fin de ayudar a alcanzar los objetivos planteados de la PR2, se ha desarrollado esta guía. Debe servir como ejemplo de cómo realizar algunas de las tareas anteriormente descritas para llevar a cabo el diseño, y el desarrollo de los procesos ETL y la carga efectiva del almacén de datos.

## 2. Identificación de los procesos ETL

A la hora de diseñar los procesos de carga de una base de datos analítica existen varias estrategias. Es habitual estructurar los procesos ETL sobre la base de las entidades de datos que deben actualizarse, ya que existen diferencias conceptuales en la actualización de una dimensión con respecto a la de una tabla de hechos. La división del proceso de carga inicial en diferentes bloques de actualización facilitará el diseño de un orden de ejecución y la gestión de las dependencias. Cada uno de estos bloques de actualización se dividirá en las correspondientes etapas de extracción, transformación y carga.

Se identifican los dos bloques siguientes:

- **Bloque IN:** procesos de carga de los datos desde las fuentes a las tablas intermedias en el área de maniobras (STG). Estos procesos se distinguen por el prefijo «IN\_» en el nombre.
- **Bloque TR:** procesos de transformación para cargar los datos desde las tablas intermedias hasta las tablas del almacén de datos, según el modelo multidimensional diseñado. Así pues, son diferentes los procesos ETL de transformación para cargar las dimensiones de aquellos que se realizan para cargar las tablas de hechos. Estos procesos se distinguen con el prefijo «TR\_» en el nombre.

## Bloque IN (de las fuentes a las tablas intermedias)

Nombre ETL	Descripción	Origen de datos	Tabla destino (stage)
IN_TAXI_ZONES	Carga la información de las ubicaciones de recogida o devolución de pasajeros en los viajes de taxis en la ciudad de Nueva York.	taxi_zone_lookup.csv	STG_TAXI_ZONES
...	...	...	...

## Bloque TR (de las tablas intermedias a las tablas del almacén de datos)

El bloque TR de procesos ETL para poblar el modelo multidimensional del almacén tiene dos partes diferenciadas. Por un lado, los procesos de carga y transformación de las dimensiones y, por otro, los de las tablas de hechos. El orden de ejecución es importante para que la carga de datos sea la correcta.

Las dimensiones se cargarán primero y, después, las tablas de hechos para que no haya errores durante la carga.

Por una parte, algunos de los procesos del bloque TR de carga y transformación de las dimensiones son los siguientes:

Nombre ETL	Descripción	Origen de datos	Tabla destino
TR_DIM_LOCATION	Carga de la dimensión con información de las zonas de NY.	STG_TAXI_ZONES	DIM_LOCATION
...	...	...	...

Por otra parte, el proceso del bloque de carga y transformación de la tabla de hechos es el siguiente:

Nombre ETL	Descripción	Tabla destino
TR_FACT_FHV_TRIP	Carga de la información para el análisis de los viajes en vehículos de alquiler.	FACT_FHV_TRIP
...	...	...

**Nota:** en este punto, el estudiantado deberá completar la identificación de los procesos de cada uno de los bloques (IN y TR) que se desarrollarán para cargar las dimensiones, así como las tablas de hechos del modelo multidimensional del almacén de datos.

## 3. Diseño y desarrollo de los procesos ETL (60 %)

En este apartado, se deben diseñar los procesos de carga identificados en el punto anterior con la herramienta de diseño proporcionada. En este caso es Pentaho Data Integration (PDI).

### 3.1. Creación de tablas

El primer paso para la implementación de los procesos ETL consiste en crear las tablas. Esto se llevará a cabo una única vez, mediante *scripts*, sobre la base de datos proporcionada (en nuestro caso, SQL Server). Se deberán crear las tablas intermedias y las tablas del modelo dimensional de la solución oficial; es decir, las dimensiones y las tablas de hechos. Para hacerlo, debe utilizarse el diseño físico propuesto en la solución de la PR1.

Una vez implementado el modelo físico del almacén, el siguiente paso es el diseño de los procesos ETL de cada uno de los bloques (IN y TR). Estos procesos permitirán poblar las tablas del área intermedia (STG), las de dimensiones y las de hechos del almacén de datos que se ha diseñado.

A continuación, se muestra un ejemplo de creación para las tablas intermedias:

#### Tabla intermedia «STG\_TAXI\_ZONES»

```
CREATE TABLE STG_TAXI_ZONES
(
    LocationID INT
    , Borough VARCHAR(50)
    , "Zone" VARCHAR(100)
    , Service_zone VARCHAR(100)
)
;
```

Ejemplo de una dimensión:

#### Tabla explotación « DIM\_LOCATION »

```
CREATE TABLE [dbo].[DIM_LOCATION](
    locationID INT
    , borough VARCHAR(50)
    , "zone" VARCHAR(100)
    , service_zone VARCHAR(100)
    CONSTRAINT [PK_ DIM_LOCATION] PRIMARY KEY CLUSTERED)
```

## 3.2. Establecer el entorno

Es una buena práctica utilizar variables de entorno para evitar introducir errores en definiciones repetitivas durante la implementación de los procesos. PDI permite añadir variables personalizadas y propias de los desarrollos en el archivo «Kettle.properties».

En el caso tratado, se pueden utilizar tres variables:

- Una para almacenar la ruta de las fuentes de datos «DIR\_ENT».
- Una cadena de conexión a la base de datos «CN\_STAGE» (área intermedia o *staging area*).
- Una cadena de conexión al almacén de datos «CN\_DW» (*data warehouse*).

Se puede crear un esquema *stage* en el SQL Server dentro de la base de datos que se tenga asignada para cargar las tablas intermedias («IN\_») y definir la variable «CN\_STAGE» haciendo referencia a este esquema, pero para simplificar la solución de la práctica se cargarán todas las tablas al esquema por defecto, *dbo*.

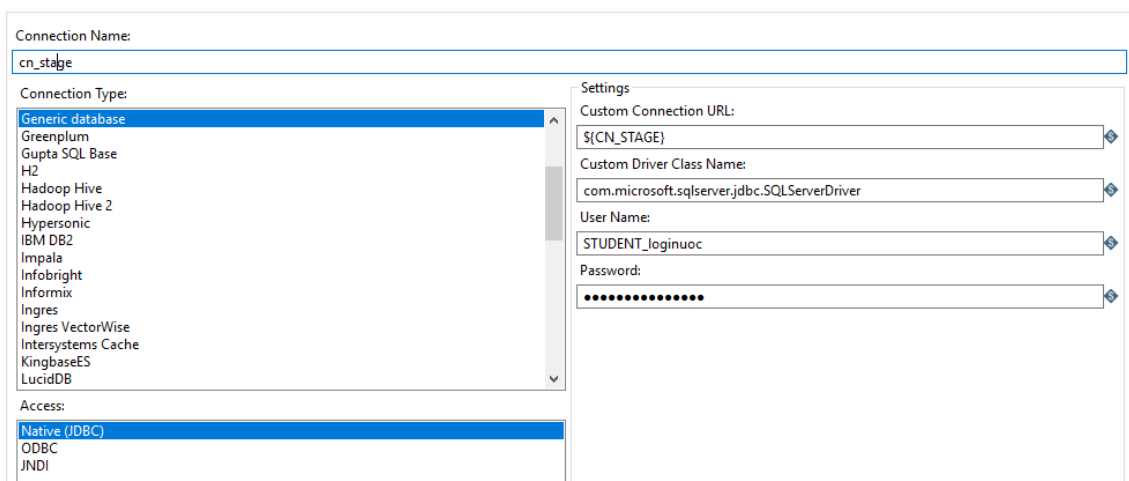
Variable	Valor
DIR_ENT	F:\fuentes
CN_STAGE	jdbc:sqlserver://UCS1R1UOCSQLxx:1433;databaseName=DB_loginuoc;integratedSecurity=false
CN_DW	jdbc:sqlserver://UCS1R1UOCSQLxx:1433;databaseName=

## 3.3. Conexiones

Otro paso previo que se debe realizar es crear las conexiones a las bases de datos que se usan en todas las transformaciones y trabajos de los procesos de carga.

Se han definido dos conexiones diferentes, una para la BD del modelo multidimensional y otra para el área intermedia («STAGE»); de esta manera se diferenciarán claramente su uso, aunque físicamente se refieran al mismo esquema de la BD.

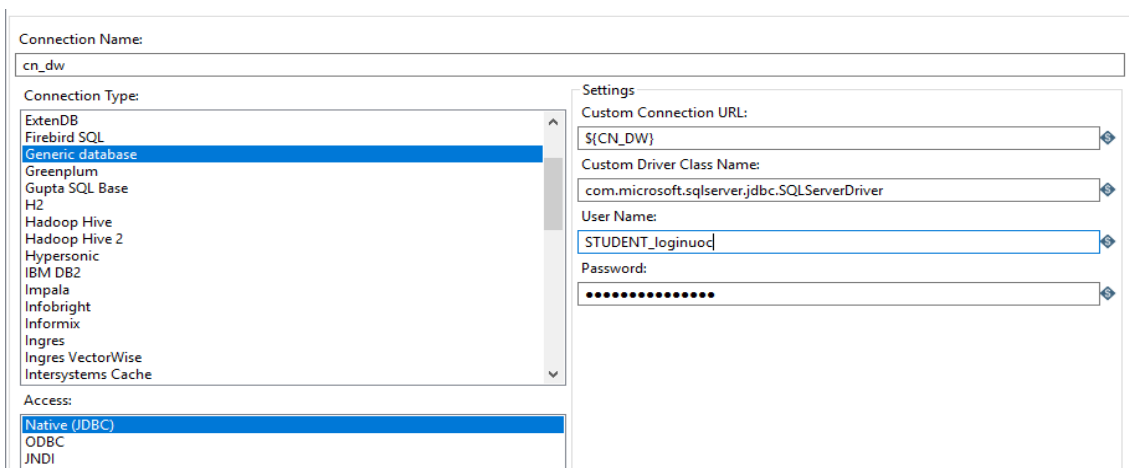
Para crear una nueva conexión hay que pulsar sobre «File» > «New» > «Database Connection». En la creación de la conexión al «STAGE», el nombre que se les asignará es «cn\_stage»:



The screenshot shows the 'Database Connection' dialog box for a connection named 'cn\_stage'. The 'Connection Name' field is filled with 'cn\_stage'. The 'Connection Type' list on the left has 'Generic database' selected. The 'Access' list at the bottom has 'Native (JDBC)' selected. The 'Settings' panel on the right contains the following fields:

- Custom Connection URL:** \${CN\_STAGE}
- Custom Driver Class Name:** com.microsoft.sqlserver.jdbc.SQLServerDriver
- User Name:** STUDENT\_loginuoc
- Password:** (masked with dots)

En la creación de la conexión al «DW», el nombre que se le dará es «cn\_dw»:



The screenshot shows the 'Database Connection' dialog box for a connection named 'cn\_dw'. The 'Connection Name' field is filled with 'cn\_dw'. The 'Connection Type' list on the left has 'Generic database' selected. The 'Access' list at the bottom has 'Native (JDBC)' selected. The 'Settings' panel on the right contains the following fields:

- Custom Connection URL:** \${CN\_DW}
- Custom Driver Class Name:** com.microsoft.sqlserver.jdbc.SQLServerDriver
- User Name:** STUDENT\_loginuoc
- Password:** (masked with dots)

## 3.4. Bloque IN

### Transformación de «IN\_TAXI\_ZONES»

A continuación, se describe parte del desarrollo de la transformación de «IN\_TAXI\_ZONES» (identificada en el primer punto de la guía) mediante Spoon. El objetivo es cargar uno de los orígenes de los datos identificados, «taxi\_zone\_lookup.csv», en la tabla «STG\_TAXI\_ZONES» del área intermedia (*staging area*). Dicha tabla intermedia, cuyo *script* se habrá escrito en el apartado «Creación», se deberá crear con anterioridad en la base de datos analítica.

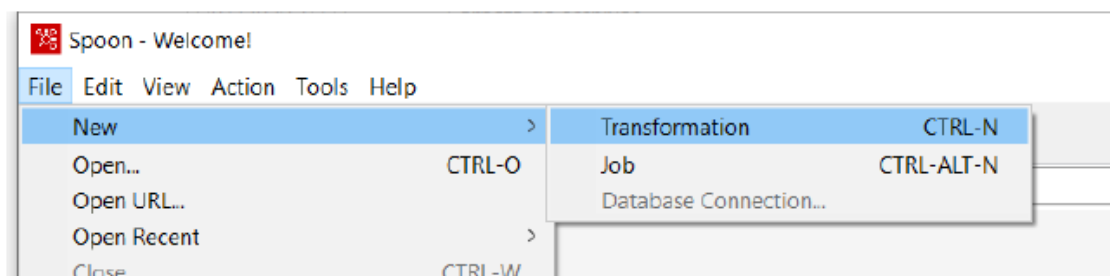
Para este caso práctico se han utilizado fuentes externas (no operacionales) que se emplearán para descubrir el conocimiento mediante el análisis de los datos. Es muy habitual manipular los ficheros realizando manualmente una serie de acciones de preparación antes de su procesamiento (preprocesado).

La transformación de «IN\_TAXI\_ZONES» contiene las siguientes alteraciones:

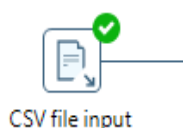
- Lectura del fichero .csv,
- Carga a la tabla intermedia «STG\_TAXI\_ZONES».

A continuación, se detallan las diferentes etapas que se deben implementar para realizar la carga de datos.

En primer lugar, se crea una nueva transformación:



En ella, como primer paso, se debe realizar la entrada del fichero .csv. Para ello hay que utilizar el tipo «CSV file input». En este paso se debe indicar el fichero desde el cual se extraen los datos. Para ello se debe utilizar la variable de entorno «DIR\_IN» creada previamente e indicar el tipo de motor que hay que usar.



Para realizar correctamente la carga, se deberán de indicar los parámetros correctamente, destacando en este caso lo siguiente:

- Indicar como delimitador la coma (`,`).
- Desmarcar «Lazy conversion?».



CSV file input

Step name: S

Filename: C:\taxi\Fonts\taxi\_zone\_lookup.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? ☐

Format: mixed

File encoding:

#	Name	Type	Format	Length	Precision	Currency	Decimal	Group	Trim type
1	LocationID	Integer	#	15	0	\$	,	.	none
2	Borough	String		50		\$	,	.	none
3	Zone	String		100		\$	,	.	none
4	Service_zone	String		100		\$	,	.	none

Help OK Get Fields Preview Cancel

**Nota:** en este paso, también se deberá de indicar los campos que se van a tratar con el botón «Get fields» y se completará su definición. Hay que especificar, donde se considere necesario, la precisión y la longitud de los campos.

Se puede realizar una visualización previa de los datos que se cargarán con el botón «Previsualizar filas».

Examine preview data

Rows of step: S (265 rows)

#	LocationID	Borough	Zone	Service_zone
1	1	EWB	Newark Airport	EWB
2	2	Queens	Jamaica Bay	Boro Zone
3	3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	4	Manhattan	Alphabet City	Yellow Zone
5	5	Staten Island	Arden Heights	Boro Zone
6	6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	7	Queens	Astoria	Boro Zone
8	8	Queens	Astoria Park	Boro Zone
9	9	Queens	Auburndale	Boro Zone
10	10	Queens	Baisley Park	Boro Zone
11	11	Brooklyn	Bath Beach	Boro Zone
12	12	Manhattan	Battery Park	Yellow Zone
13	13	Manhattan	Battery Park City	Yellow Zone
14	14	Brooklyn	Bay Ridge	Boro Zone
15	15	Queens	Bay Terrace/Fort Totten	Boro Zone
16	16	Queens	Bayside	Boro Zone
17	17	Brooklyn	Bedford	Boro Zone
18	18	Bronx	Bedford Park	Boro Zone
19	19	Queens	Bellerose	Boro Zone
20	20	Bronx	Belmont	Boro Zone
21	21	Brooklyn	Bensonhurst East	Boro Zone
22	22	Brooklyn	Bensonhurst West	Boro Zone
23	23	Staten Island	Bloomfield/Emerson Hill	Boro Zone
24	24	Manhattan	Bloomingdale	Yellow Zone
25	25	Brooklyn	Boerum Hill	Boro Zone
26	26	Brooklyn	Borough Park	Boro Zone
27	27	Queens	Breezy Point/Fort Tilden/Riis Beach	Boro Zone
28	28	Queens	Briarwood/Jamaica Hills	Boro Zone

Y, por último, se cargarán los datos en la tabla intermedia del *stage*, utilizando el paso «Table output». Para este paso se puede reutilizar la conexión previamente creada como «cn\_stage».

El paso de cargar los datos a la tabla intermedia del *stage* se debe configurar como se indica en el menú principal que muestra a continuación. Es importante tener la tabla de STG creada para insertar los datos preprocesados.

Como se puede ver en la imagen, para dejar la transformación preparada para posibles reprocesos, es necesario realizar un borrado previo para actualizar los datos en el caso de que se tuviera que efectuar. Para esto, se activará el *check* «Truncate table», situado en los campos de la base de datos.

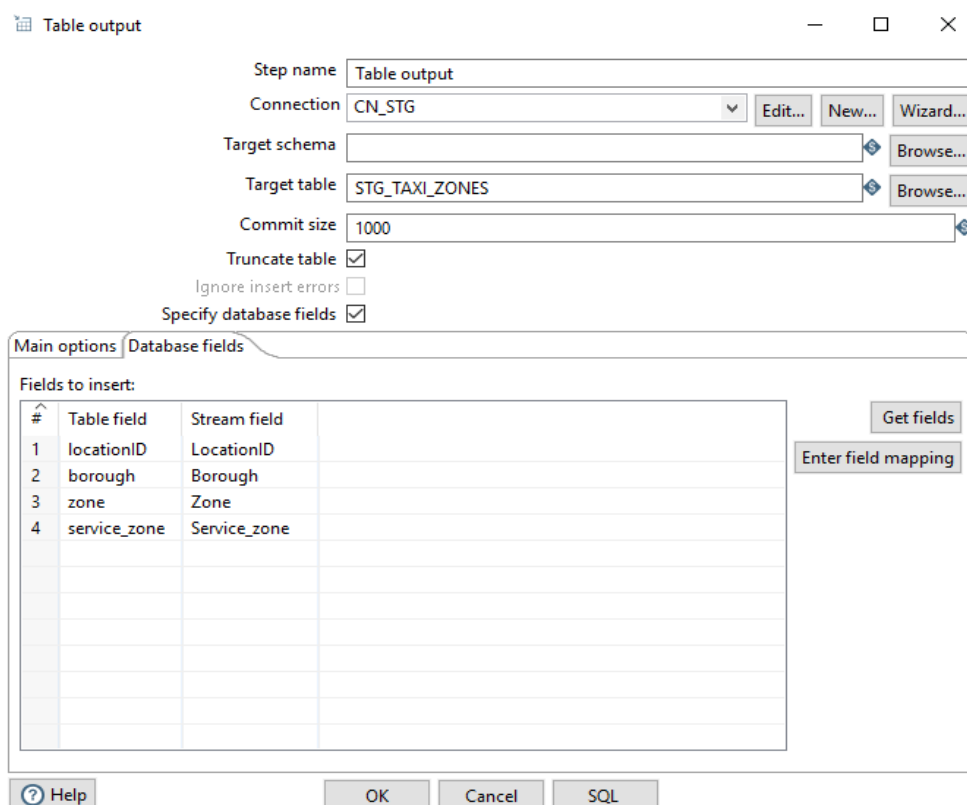


Table output

Step name: Table output

Connection: CN\_STG [Edit...] [New...] [Wizard...]

Target schema: [Browse...]

Target table: STG\_TAXI\_ZONES [Browse...]

Commit size: 1000

Truncate table: ☒

Ignore insert errors: ☐

Specify database fields: ☒

Main options | Database fields

Fields to insert:

#	Table field	Stream field	
1	locationID	LocationID	
2	borough	Borough	
3	zone	Zone	
4	service_zone	Service_zone	

[Get fields] [Enter field mapping]

[Help] [OK] [Cancel] [SQL]

El proceso de la transformación completa es el siguiente:

**Execution Results**

Logging Execution History Step Metrics Performance Graph Metrics Preview data

2024/04/18 19:13:00 - Spoon - Running transformation using the Kettle execution engine  
 2024/04/18 19:13:00 - Spoon - Transformation opened.  
 2024/04/18 19:13:00 - Spoon - Launching transformation [IN\_TAXI\_ZONES]...  
 2024/04/18 19:13:00 - Spoon - Started the transformation execution.  
 2024/04/18 19:13:00 - IN\_TAXI\_ZONES - Dispatching started for transformation [IN\_TAXI\_ZONES]  
 2024/04/18 19:13:00 - STG\_TAXI\_ZONES.0 - Connected to database [CN\_STG] (commit=1000)  
 2024/04/18 19:13:00 - TAXI\_ZONES.0 - Header row skipped in file 'C:\TLC\FonTS\taxi\_zone\_lookup.csv'  
 2024/04/18 19:13:00 - TAXI\_ZONES.0 - Finished processing (I=266, O=0, R=0, W=265, U=0, E=0)  
 2024/04/18 19:13:00 - STG\_TAXI\_ZONES.0 - Finished processing (I=0, O=265, R=265, W=265, U=0, E=0)  
 2024/04/18 19:13:00 - Spoon - The transformation has finished!!

Y se puede comprobar en la base de datos el resultado del proceso realizado:

```

1
2 SELECT [LocationID]
3      , [Borough]
4      , [Zone]
5      , [Service_zone]
6 FROM [dbo].[STG_TAXI_ZONES]
  
```

100 %

Results Messages

	LocationID	Borough	Zone	Service_zone
1	1	EWB	Newark Airport	EWB
2	2	Queens	Jamaica Bay	Boro Zone
3	3	Bronx	Allerton/Pelham Gardens	Boro Zone
4	4	Manhattan	Alphabet City	Yellow Zone
5	5	Staten Island	Arden Heights	Boro Zone
6	6	Staten Island	Arrochar/Fort Wadsworth	Boro Zone
7	7	Queens	Astoria	Boro Zone
8	8	Queens	Astoria Park	Boro Zone
9	9	Queens	Aubumdale	Boro Zone
10	10	Queens	Baisley Park	Boro Zone
11	11	Brooklyn	Bath Beach	Boro Zone
12	12	Manhattan	Battery Park	Yellow Zone
13	13	Manhattan	Battery Park City	Yellow Zone
14	14	Brooklyn	Bay Ridge	Boro Zone
15	15	Queens	Bay Terrace/Fort Totten	Boro Zone
16	16	Queens	Bayside	Boro Zone
17	17	Brooklyn	Bedford	Boro Zone

**Nota:** para la entrega de la PR2, el estudiantado deberá diseñar todos los procesos ETL de cada uno de los bloques (IN y TR) individualmente. En este ejemplo se ha mostrado un caso básico de carga de datos, pero, según el formato de origen de los datos y de la calidad de estos, tal vez sea necesario utilizar otras transformaciones. Spoon dispone de una gran cantidad de componentes, organizados por categorías, a los que se puede acceder desde el menú lateral.

## 4. Implementación de los trabajos con procesos ETL (20%)

Los bloques de procesos ETL implementados que hay que tener en cuenta son los siguientes:

- «**Bloque IN**»: procesos ETL de transformación y carga al área intermedia.
- «**Bloque TR\_DIM**»: procesos ETL de transformación y carga de dimensiones.
- «**Bloque TR\_FACT**»: procesos ETL de transformación y carga de hechos.

En este punto, para realizar la carga efectiva de los datos, el estudiantado, de forma individual, debe diseñar mediante PDI los trabajos (*jobs*) que permitan la ejecución de todos los procesos ETL incluidos en cada bloque. En este apartado se deben acreditar que se han procesado los registros con una captura de pantalla de la pestaña «Logging» (se observa el número de registros cargados en cada tabla) o «Job metrics», similar a la que se muestra a continuación:

The screenshot shows the Spoon PDI interface. At the top, there's a toolbar with icons for running, saving, and other actions. Below the toolbar, a job flow is visible with three main components: 'Start', 'IN\_TAXI\_ZONES', and 'Success'. Each component has a green checkmark icon above it, indicating successful execution. Below the job flow, the 'Execution Results' tab is selected, showing a table of job metrics.

Job / Job Entry	Comment	Result
Job: JOB_IN	Start of job execution	
Start	Start of job execution	
Start	Job execution finished	Success
IN_TAXI_ZONES	Start of job execution	
IN_TAXI_ZONES	Job execution finished	Success
Success	Start of job execution	
Success	Job execution finished	Success
Job: JOB_IN	Job execution finished	Success

Adicionalmente, para acreditar la carga efectiva de los registros en la base de datos, se incorporará una captura de pantalla de SSMS con un «SELECT \* FROM TaulaQueAcabemDeCarregar» y los primeros registros obtenidos.

## 5. Formato y fecha de entrega

La entrega de esta actividad debe realizarse a través del enlace «Entrega PR2» del aula, enviando un único archivo en formato Word o PDF. El nombre del archivo debe ser la composición del nombre de usuario y «\_BDA\_PR2».

Por ejemplo, si el nombre de usuario es «bantich», el nombre del archivo debe ser «bantich\_BDA\_PR2.pdf».

**La fecha máxima de entrega es el **XX/XX/XXXX** a las 23:59 horas.**