

# Trabajo 5

## Manejo de contenido basado en URLs

Facultad De Ingeniería, Universidad De Cuenca

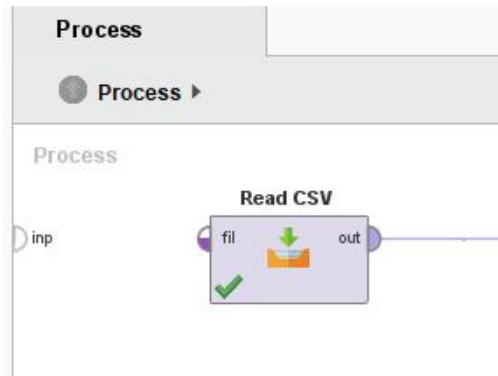
TEXT MINING

Freddy L. Abad L.

[ffreddy.abadl@ucuenca.edu.ec](mailto:ffreddy.abadl@ucuenca.edu.ec)

### PASO A

Lectura de Archivo del resultado en práctica 4, con los links de los discursos presidenciales (el proceso de la práctica 4, mostró como resultado un archivo CSV y no xlsx como se propone, no tiene mayor incidencia mas que la comodidad de uso de archivos).



**Botón Import Configuration Wizard:** Configuración de Archivo CSV, en el cual los separadores son ',' y las comillas ""

La imagen muestra la ventana de configuración 'Parameters' para el proceso 'Read CSV'. En la parte superior, hay un botón 'Import Configuration Wizard...'. A continuación, se listan varios parámetros de configuración:

- csv file:** TM\Salida.csv
- column separators:** ,
- trim lines:** ☐
- use quotes:** ☒
- quotes character:** "
- escape character:** \
- skip comments:** ☒
- comment characters:** #
- starting row:** 1
- parse numbers:** ☒

En la parte inferior, hay un botón 'Hide advanced parameters' y un enlace 'Change compatibility (9.2.001)'.


## PASO B:

Uso de Get\_pages, para el manejo de todos los links del archivo Salida.csv (Contiene el link con los discursos de todos los presidentes de EEUU, archivo resultado de la Práctica 4).



## Configuración de Parámetros de Get\_Pages

**Parameters** ✕

 **Get Pages**

link attribute

ENLACE=url2

ⓘ

page attribute

ⓘ

☐ random user agent

ⓘ

user agent

ⓘ

connection timeout

10000

ⓘ

read timeout

10000

ⓘ

☒ follow redirects

ⓘ

accept cookies

none

ⓘ

request method

GET



ⓘ

delay

none

ⓘ

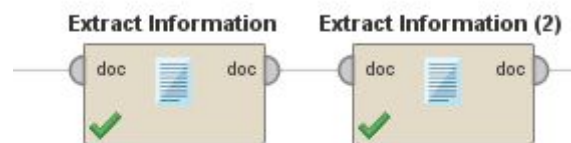
## Resultados de Get\_pages

| ExampleSet (Read CSV) <span>✕</span>  |  |  |
|---|--|--|
| Open in <span> Turbo Prep</span> <span> Auto Model</span> |  |  |
| Filter (58 / 58 examples): <span>all</span> <span>▼</span>  |  |  |
| Row No.   | ENLACE=url2  | gensym1  |
| 1   | https://www.presidency.ucsb.edu/ws/index.php?pid=25800 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 2   | https://www.presidency.ucsb.edu/ws/index.php?pid=25801 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 3   | https://www.presidency.ucsb.edu/ws/index.php?pid=25802 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 4   | https://www.presidency.ucsb.edu/ws/index.php?pid=25803 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 5   | https://www.presidency.ucsb.edu/ws/index.php?pid=25804 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 6   | https://www.presidency.ucsb.edu/ws/index.php?pid=25805 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 7   | https://www.presidency.ucsb.edu/ws/index.php?pid=25806 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 8   | https://www.presidency.ucsb.edu/ws/index.php?pid=25807 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 9   | https://www.presidency.ucsb.edu/ws/index.php?pid=25808 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 10  | https://www.presidency.ucsb.edu/ws/index.php?pid=25809 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 11  | https://www.presidency.ucsb.edu/ws/index.php?pid=25810 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 12  | https://www.presidency.ucsb.edu/ws/index.php?pid=25811 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 13  | https://www.presidency.ucsb.edu/ws/index.php?pid=25812 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |
| 14  | https://www.presidency.ucsb.edu/ws/index.php?pid=25813 | <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML+RDFa 1.0//EN" |

## PASO C



Ingresamos al proceso “Process Documents”, se agrega la herramienta **Extract Information**, por medio de la cual manejaremos expresiones regulares, con el fin de manejar Nombre del Presidente, Año de Discurso Inaugural y Discurso Presidencial de cada Toma de Posesión (link del archivo Salida.csv)



### Expresiones Regulares del Extract Information

Para la configuración de Expresiones Regulares se debe seleccionar el botón “Edit List” en el panel de configuración de **Extract Information**.

**Parameters** X

**Extract Information**

query type: Regular Expression

attribute type: Nominal

regular expression queries: Edit List (3)...

Nombramos tres atributos con los nombres de: Fecha Inauguración, Presidente, Texto

Edit Parameter List: regular expression queries

Edit Parameter List: **regular expression queries**  
Specifies a list of attribute names and their corresponding regular expressions. The first matching group is used as value. See the operator documentation for details on regular expressions.

| attribute name | query expression                                       |
|----------------|--|
| Presidente     | <a href="/people/president/[a-z0-9-]+>([ a-zA-Z]+)</a> |
| Fecha Discurso | "([0-9]{4}-[0-9]{2}-[0-9]{2})T                         |

**Parameters** X

**Extract Information (2) (Extract Information)**

query type String Matching ▼ ⓘ

string machting queries Edit List (1)... ⓘ

attribute type Nominal ▼ ⓘ

Edit Parameter List: string machting queries X

**Edit Parameter List: string machting queries**  
Specifies a list of string matching start and end sequences. Everything between will be used as result. See the operator documentation for details on string matching.

| attribute name | query expression                        |
|----------------|---|
| Discurso       | <div class="field-docs-content"> </div> |

Para seleccionar los atributos necesarios a guardar en la base de datos se procederá a Seleccionar atributos mediante el proceso **Select Attribute**



Select Attributes: attributes X

**Select Attributes: attributes**  
The attribute which should be chosen.

Attributes

Search X

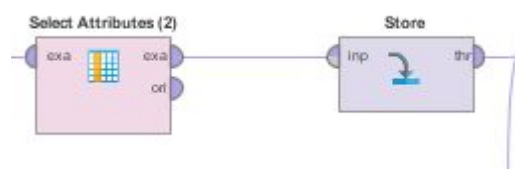
ENLACE=url2

Selected Attributes

Search + X

Discurso  
Fecha Discurso  
Presidente

A continuación se debe emplear el proceso **Repository Access:Store** para guardar información en un repositorio de RapidMiner llamado Discursos



Configuracion de Parametros de Store

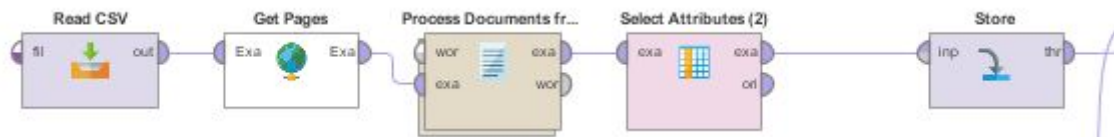
Parameters

Store

repository entry

Discursos

Resultados del proceso **Store**



ExampleSet (Select Attributes (2))

Open in

Turbo Prep

Auto Model

| Row No. | Presidente       | Fecha Discu... | Discurso       |
|---------|------------------|----------------|----------------|
| 1       | John Quincy ...  | 1825-03-04     | <p>In com...   |
| 2       | George Was...    | 1793-03-04     | <p>Fellow ...  |
| 3       | John Adams       | 1797-03-04     | <p>WHEN ...    |
| 4       | Thomas Jeffe...  | 1801-03-04     | <p>Friends...  |
| 5       | Thomas Jeffe...  | 1805-03-04     | <p>PROC...     |
| 6       | James Madis...   | 1809-03-04     | <p>Unwilli...  |
| 7       | James Madis...   | 1813-03-04     | <p>About t...  |
| 8       | James Monroe     | 1817-03-04     | <p>I shoul...  |
| 9       | James Monroe     | 1821-03-04     | <p>Fellow-...  |
| 10      | John Quincy ...  | 1825-03-04     | <p>In com...   |
| 11      | Andrew Jack...   | 1829-03-04     | <p>Fellow-...  |
| 12      | Andrew Jack...   | 1833-03-04     | <p>Fellow-...  |
| 13      | Martin van Bu... | 1837-03-04     | <p>Fellow-...  |
| 14      | William Henr...  | 1841-03-04     | <p>Called f... |
| 15      | James K. Polk    | 1845-03-04     | <p>Fellow-...  |

ExampleSet (58 examples, 0 special attributes, 3 regular attributes)

Procedemos a crear un proceso de **Repository Access: Retrieve** para acceder al repositorio almacenado Discursos.



### Configuración de Parámetros de Retrieve

**Parameters** [X]

Retrieve

repository entry

El proceso a continuación que se debe emplear es **Rename attributes** el cual permite cambiar los nombres de dos atributos. Estos cambios de nombre mejoran la legibilidad de la salida.



**Parameters** [X]

Rename (2) (Rename)

old name

new name

rename additional attributes Edit List (2)...

### Configuración de lista de parametros a renombrar.

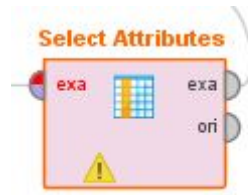
Edit Parameter List: rename additional attributes [X]

Edit Parameter List: **rename additional attributes**  
A list that can be used to define additional attributes that should be renamed.

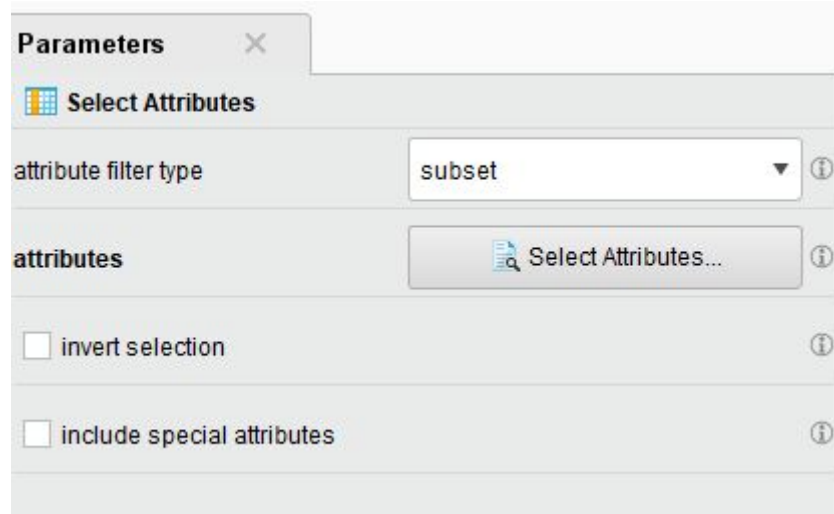
| old name       |   | new name   |
|----------------|---|------------|
| Fecha Discurso | ▼ | fecha      |
| Presidente     | ▼ | presidente |

Después de este proceso, nos devolverá un listado amplio de Atributos, para lo cual se debe filtrar los que realmente ayudarán a nuestro proceso, para tal fin se debe seleccionar el proceso: **Select Attribute**, donde seleccionamos: fecha inaugural, presidente y discurso.

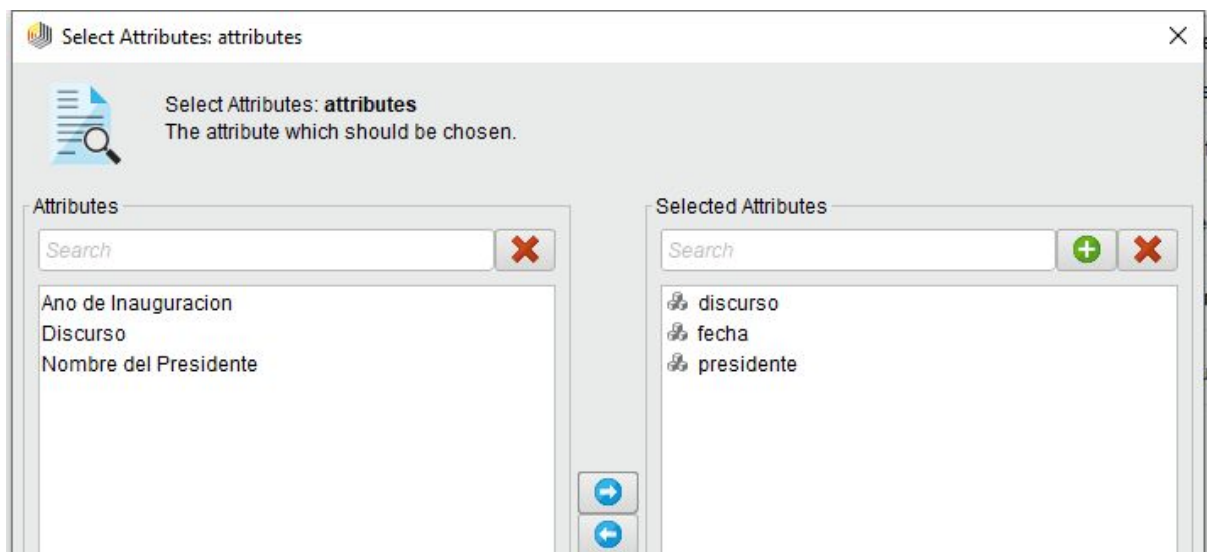




En este proceso se debe configurar qué atributos a tomar, escogemos la opción subset.



La selección de los atributos importantes para mi proceso se realiza mediante el botón “**Select Attributes**”, del cual seleccionamos Presidente, Año, Discurso.

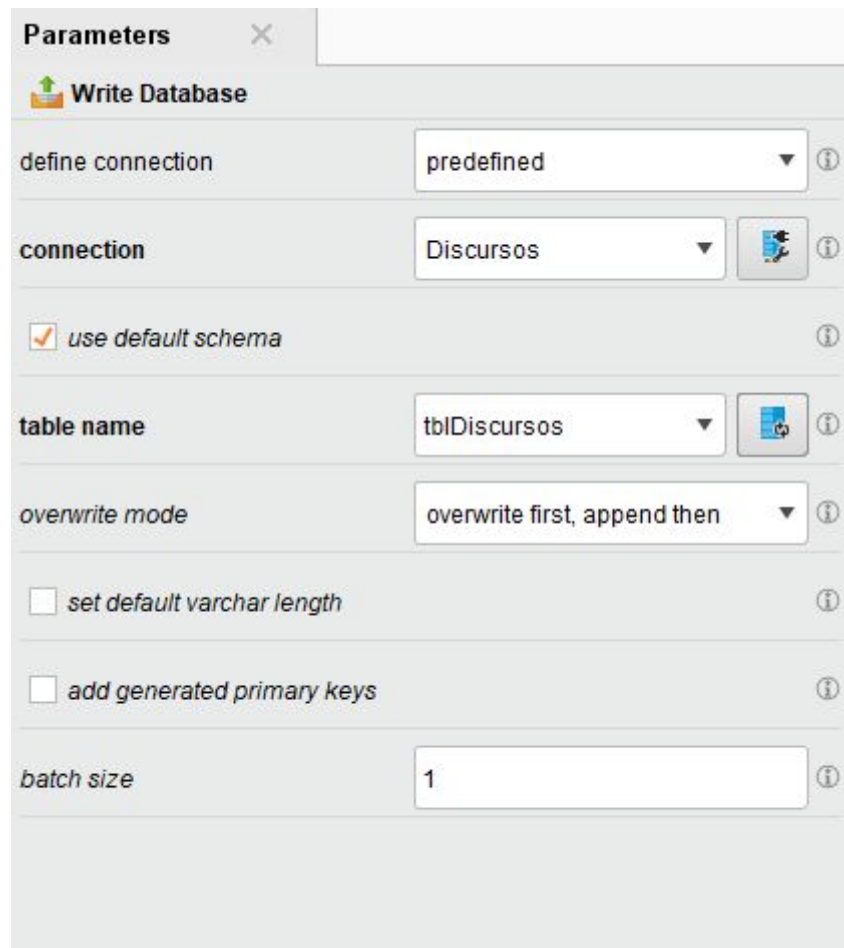


Finalmente deseamos guardar esta información en una base de datos, para lo cual seleccionamos el proceso **Write database**.



Para tal fin se debe configurar los parámetros de este proceso, se debe agregar un nombre de tabla tblDiscursos (no necesita ser creada en la base de datos del SGBD, ya que el

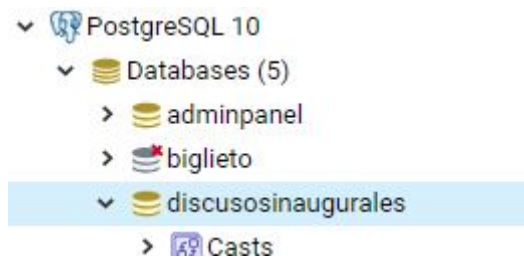
proceso de Rapidminer la creara) debemos seleccionar “*overwrite first, append then*” en la opción *overwrite mode*. Además de una conexión y configurarla



The screenshot shows the 'Parameters' dialog for the 'Write Database' operation. The parameters are as follows:

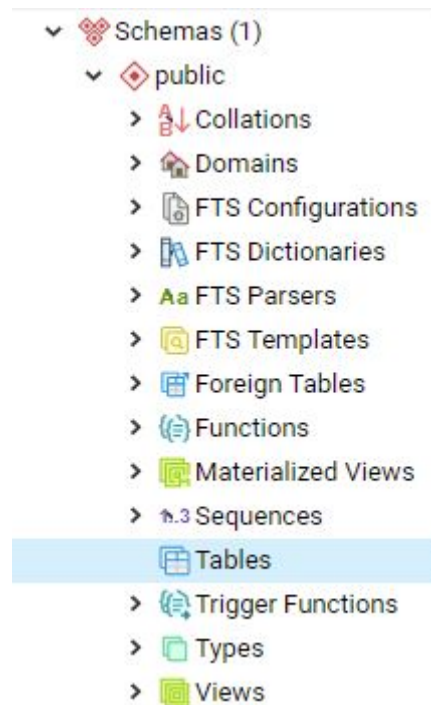
- define connection:** predefined
- connection:** Discursos
- use default schema:** ☒
- table name:** tblDiscursos
- overwrite mode:** overwrite first, append then
- set default varchar length:** ☐
- add generated primary keys:** ☐
- batch size:** 1

La conexión a la base de datos debe ser posterior a la creación de una Base de Datos(lo llamamos “discusosinaugurales”) en algún SGBD, en este caso se usó PostgreSQL como SGBD por su fácil uso, instalación, peso en disco y rapidez.

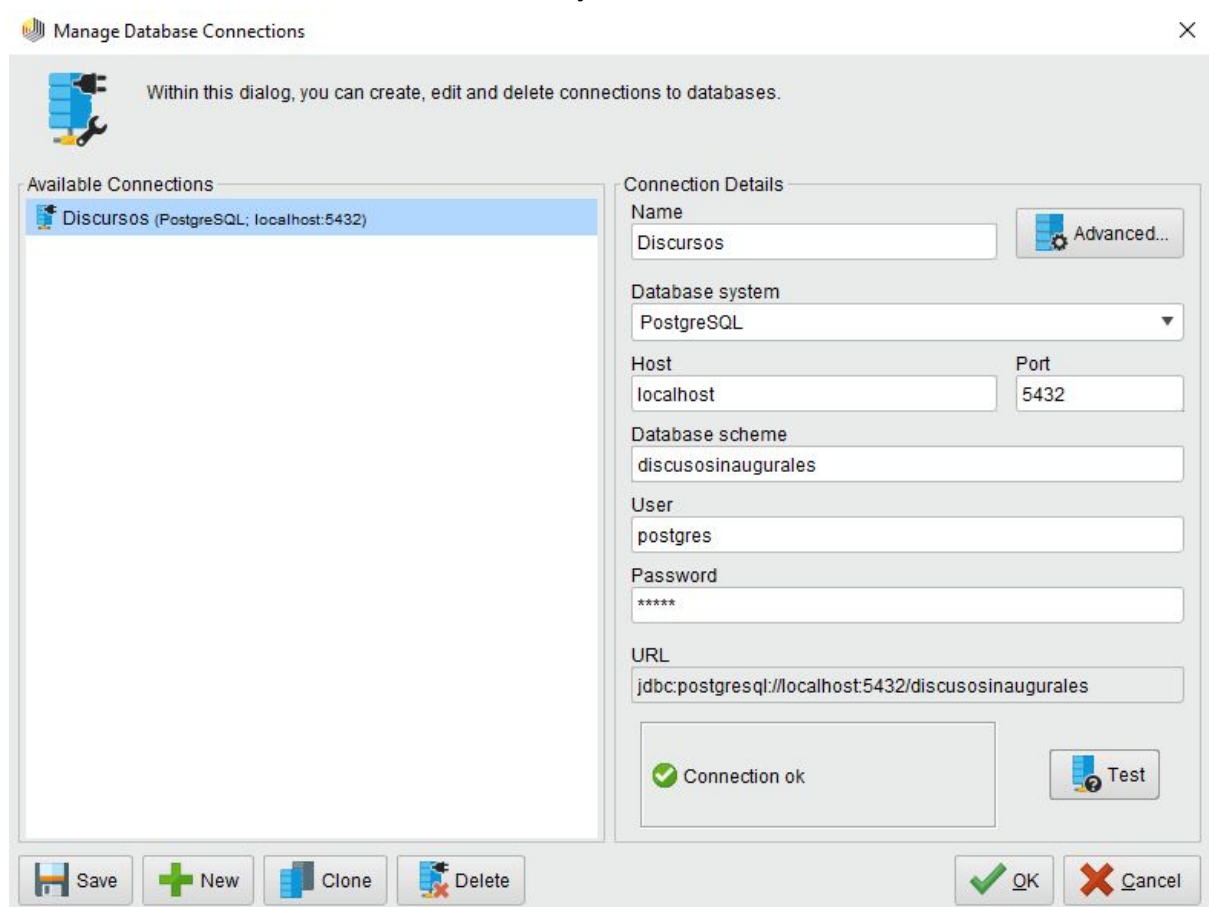


Al crear la Base de Datos no se necesita la creación de tablas, puesto que el proceso de RapidMiner las crea y maneja

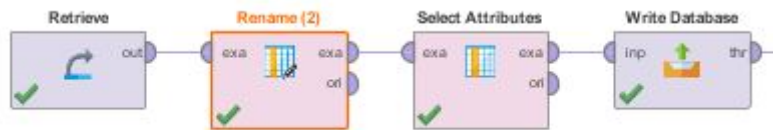




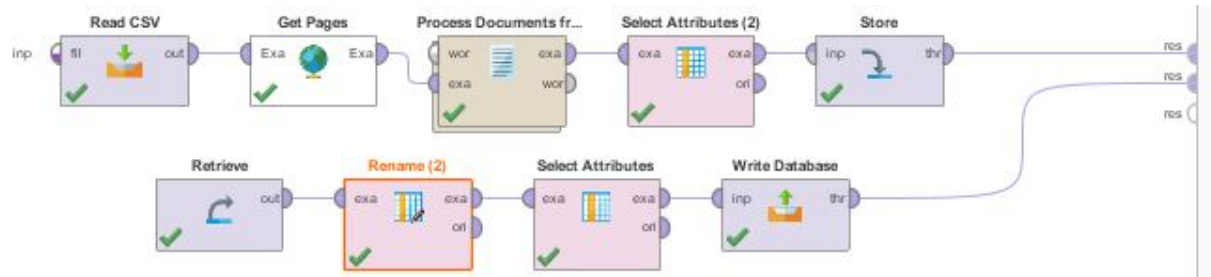
En el proceso **Write Database** se configura el host, port, username, password y finalmente se testea la conexión, obteniendo un mensaje de “Connection OK”



Resultados del proceso **Retrieve - Database**



Diseño de todo el proceso de la práctica.



ExampleSet (Select Attributes)

Examp

Open in

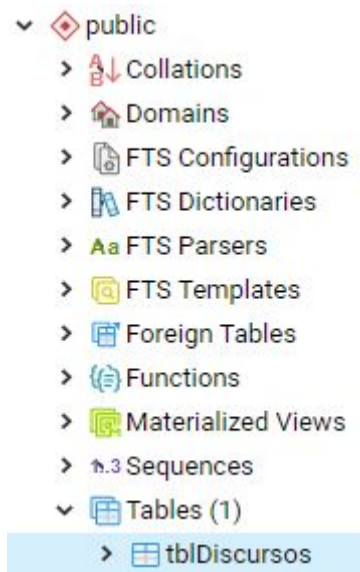
Turbo Prep

Auto Model

| Row No. | presidente       | fecha      | discurso       |
|---------|------------------|------------|----------------|
| 1       | John Quincy ...  | 1825-03-04 | <p>In com...   |
| 2       | George Was...    | 1793-03-04 | <p>Fellow ...  |
| 3       | John Adams       | 1797-03-04 | <p>WHEN ...    |
| 4       | Thomas Jeffe...  | 1801-03-04 | <p>Friends...  |
| 5       | Thomas Jeffe...  | 1805-03-04 | <p>PROC...     |
| 6       | James Madis...   | 1809-03-04 | <p>Unwilli...  |
| 7       | James Madis...   | 1813-03-04 | <p>About t...  |
| 8       | James Monroe     | 1817-03-04 | <p>I shoul...  |
| 9       | James Monroe     | 1821-03-04 | <p>Fellow-...  |
| 10      | John Quincy ...  | 1825-03-04 | <p>In com...   |
| 11      | Andrew Jack...   | 1829-03-04 | <p>Fellow-...  |
| 12      | Andrew Jack...   | 1833-03-04 | <p>Fellow-...  |
| 13      | Martin van Bu... | 1837-03-04 | <p>Fellow-...  |
| 14      | William Henr...  | 1841-03-04 | <p>Called f... |
| 15      | James K. Polk    | 1845-03-04 | <p>Fellow-...  |

ExampleSet (58 examples, 0 special attributes, 3 regular attributes)

## Verificación en base de datos PostgreSQL



|    | presidente<br>text | fecha<br>text | discurso<br>text |
|----|--------------------|---------------|------------------|
| 1  | John Quincy ...    | 1825-0...     |                  |
| 2  | George Was...      | 1793-0...     |                  |
| 3  | John Adams         | 1797-0...     |                  |
| 4  | Thomas Jeff...     | 1801-0...     |                  |
| 5  | Thomas Jeff...     | 1805-0...     |                  |
| 6  | James Madi...      | 1809-0...     |                  |
| 7  | James Madi...      | 1813-0...     |                  |
| 8  | James Monr...      | 1817-0...     |                  |
| 9  | James Monr...      | 1821-0...     |                  |
| 10 | John Quincy ...    | 1825-0...     |                  |
| 11 | Andrew Jack...     | 1829-0...     |                  |
| 12 | Andrew Jack...     | 1833-0...     |                  |