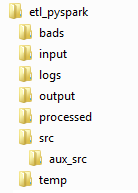
# ETL Process considerations

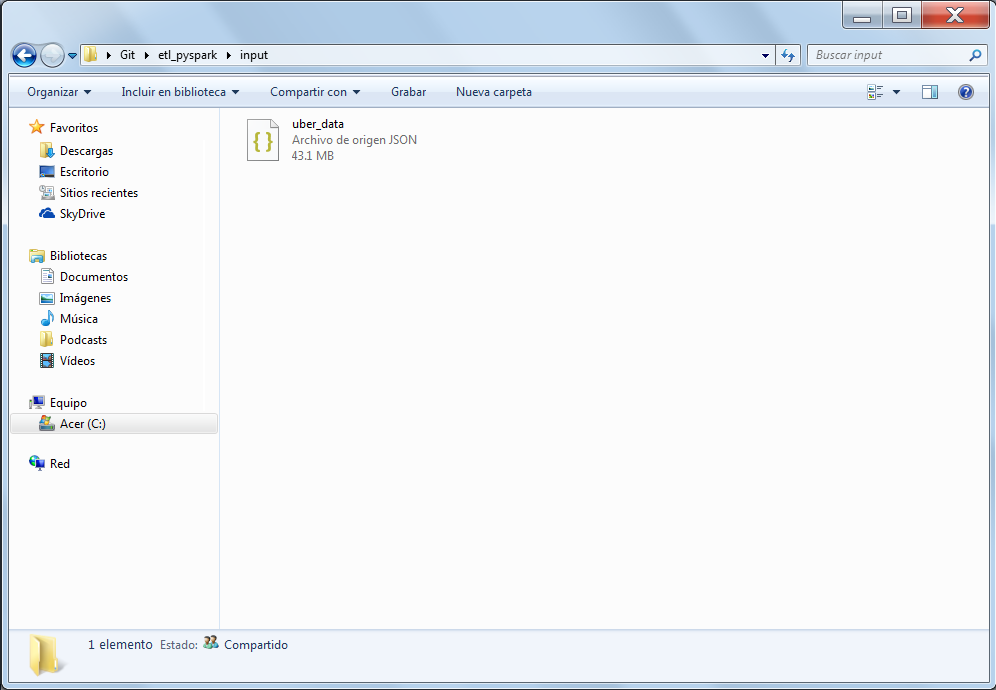
## Directories and files

For the ETL process to work correctly, the following files and directories must be present, this is after having installed all the prerequisites. In the directory where the project will be created (etl\_pyspark), the following directories must exist



### Input directory

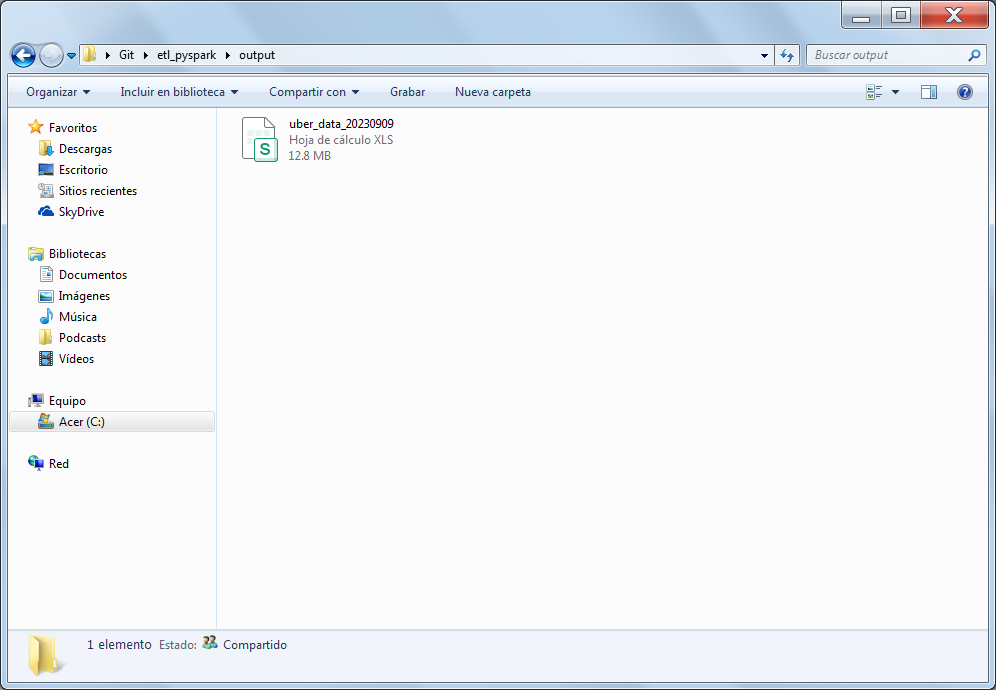
In the input directory is the uber\_data.json file that will be uploaded to the tables in the SQL server. The file used for this project is based on the file uber\_data.csv in the <https://github.com/darshilparmar/uber-etl-pipeline-data-engineering-project/tree/main/data> project.



### Output directory

The output directory is used to save all the files that are generated in each of the extract, transform and load processes. When the extraction process is run, the uber\_data\_yyyymmdd.csv file is generated with the same records as the uber\_data.json file.

Remember that the date in the name of the files created is the day the process was executed.



In the following link you can see the output log of the Extract process that creates the csv file described.



This file is the basis for creating the files in the following transform and load stages. With the exception of files that load to the following tables

CAT\_TARIFA

CAT\_TIPO\_PAGO

This is because these catalogs are created based on a file found in the following link

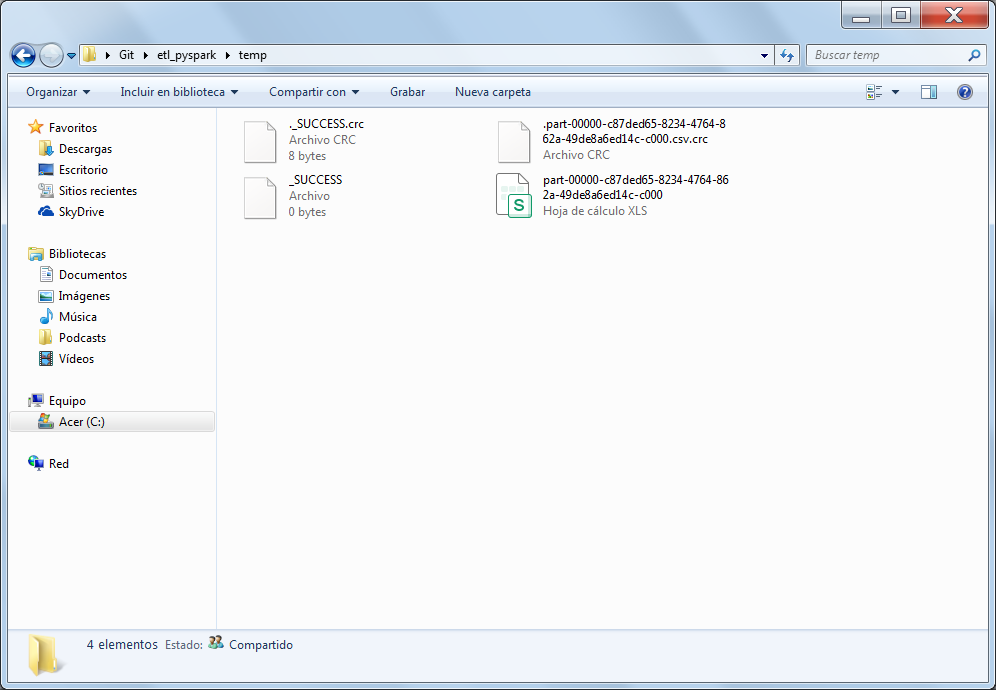
<https://www.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf>

In the case of the following table, there was no value associated with its description, so the values were invented.

CAT\_PROVEEDOR

### Temp directory

The temporary directory is used to create the csv files. This is because in pyspark, when creating a file, only the path where you want to create the file is passed, the process creates four files of which one of them is the file with the information in csv format, as you can see next.



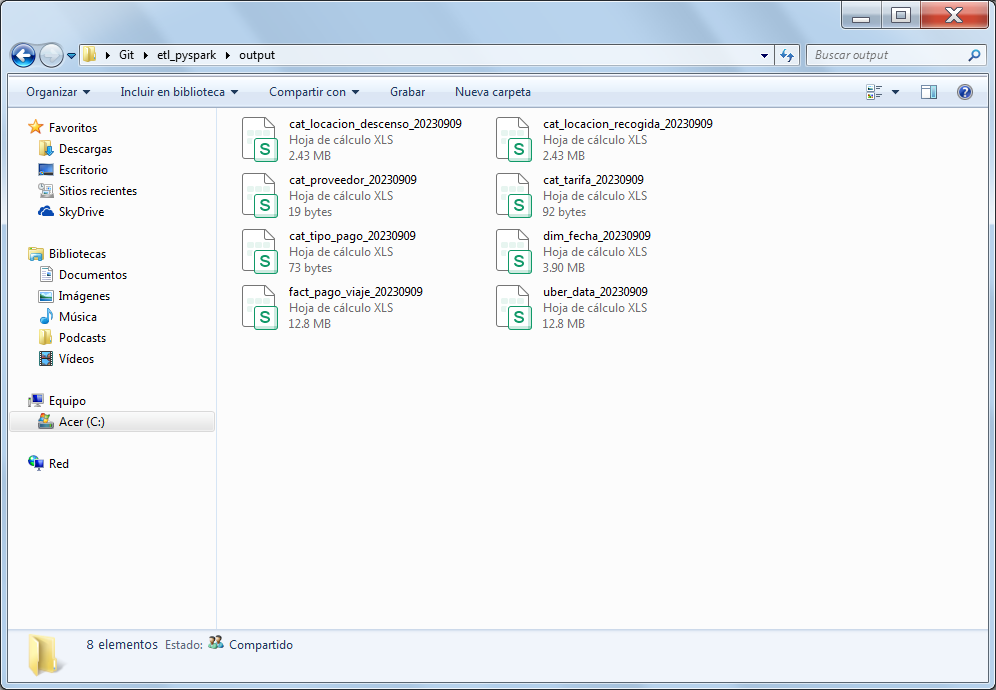
When one of the processes creates a file, the next steps are to change the name and directory of the file with the csv format to the output directory. When one of the processes creates a file, the next step is to change the name and directory of the file with the csv format to the output directory.

When the file has already been renamed and moved to the output directory, the files that are in the temporary directory are deleted, since it must always be empty for the following files that will be created.

This can be seen in the output log of the transform process, since the process generates several files.

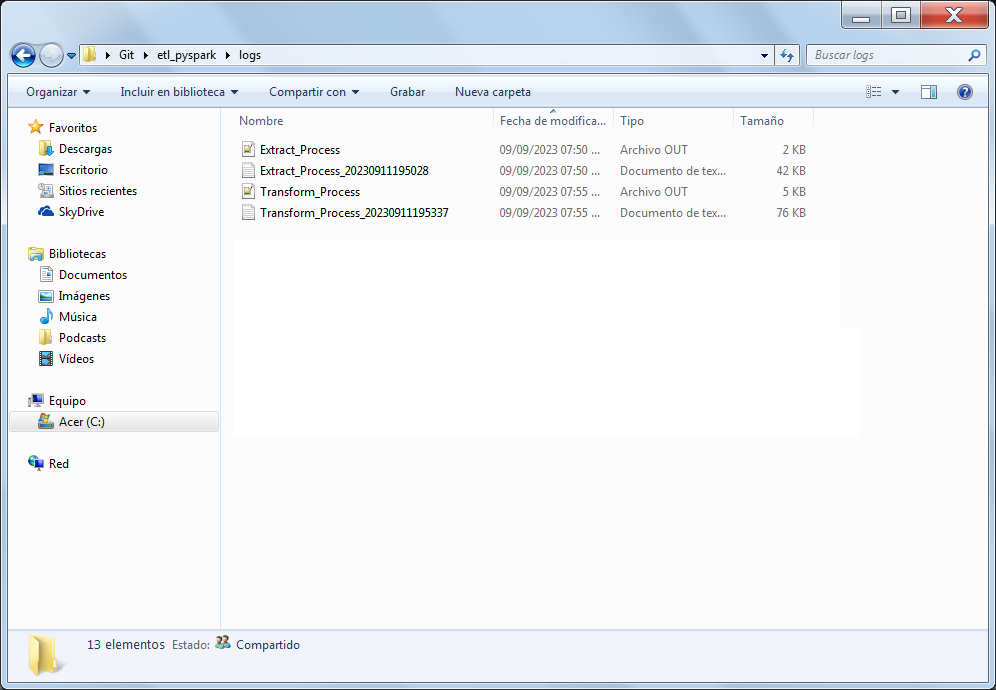


This can be seen in the output log of the transform process, since the process generates several files.



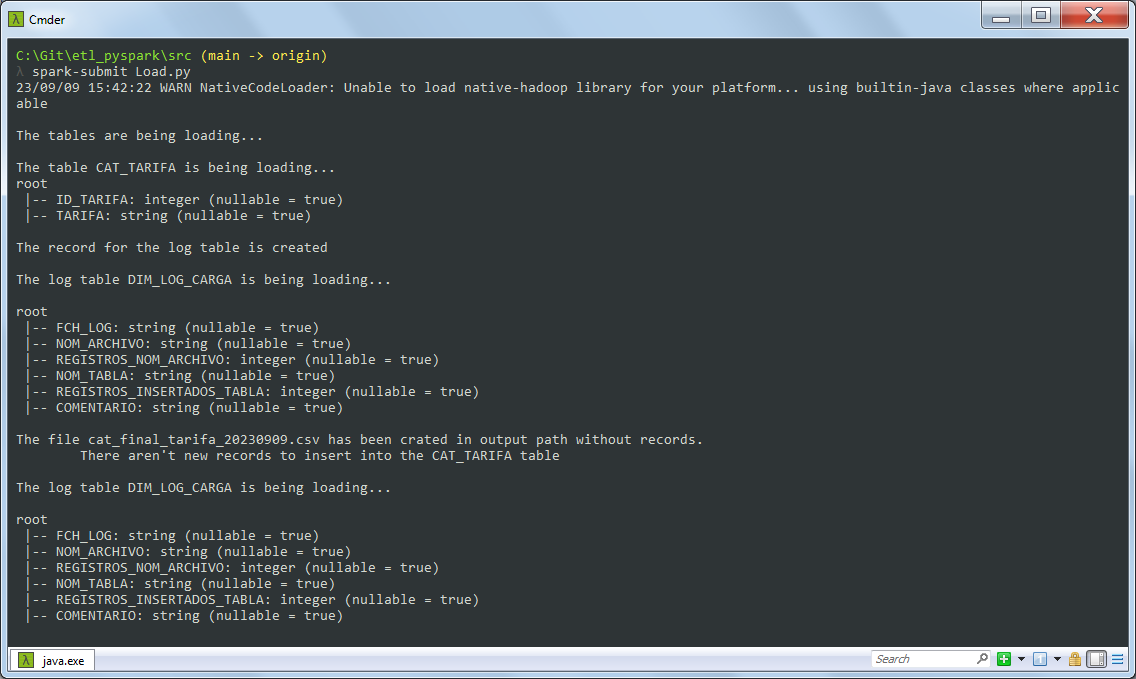
### Logs directory

The logs directory is used to create the log files. Either the files that each process generates automatically or the files with the console output logs.



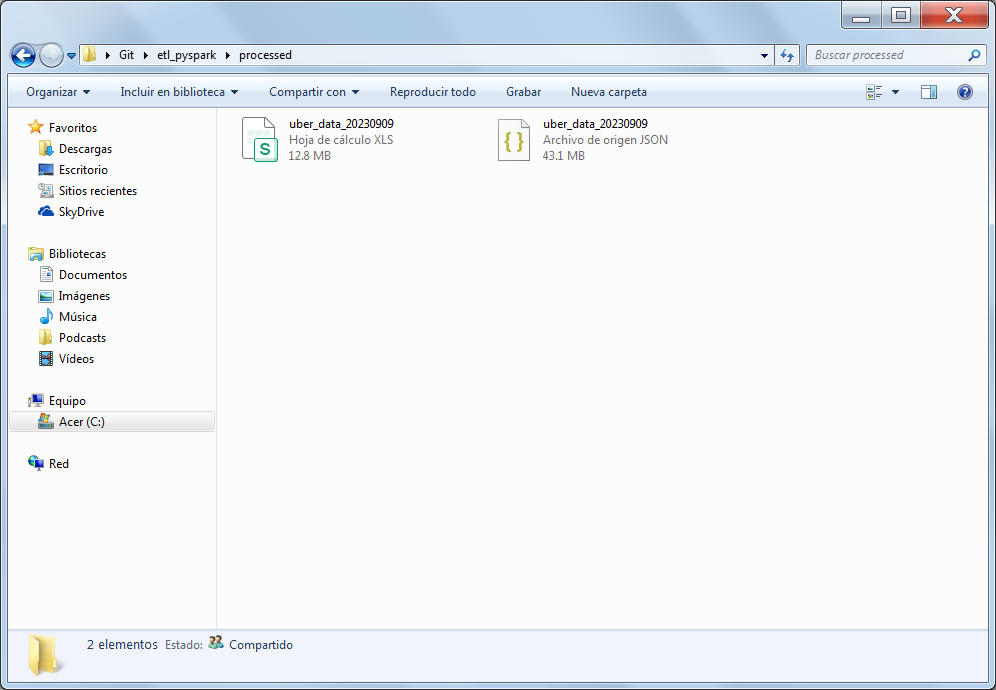
### Bads directory

The bads directory is used to create a bad file. This file is used to validate that the load to the FACT\_PAGO\_VIAJE table doesn’t have the IDs that were generated in the other tables with a null value. If everything is correct, the file should be generated without data.



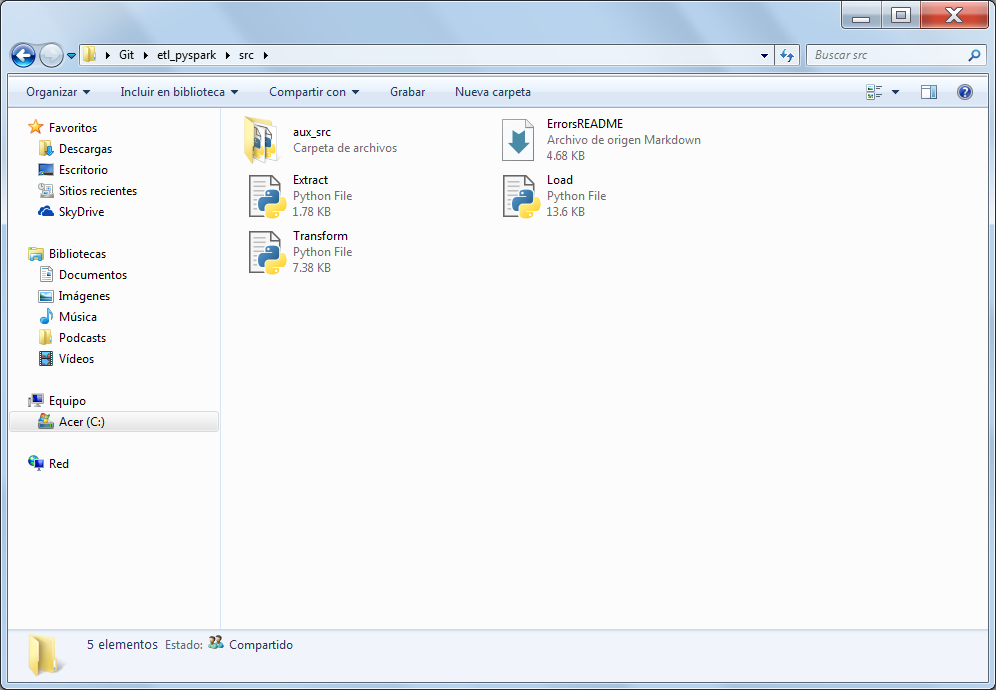
### Processed directory

The processed directory is used to copy the uber\_data.json and uber\_data\_yyyymmdd.csv files. When the tables have already been loaded.



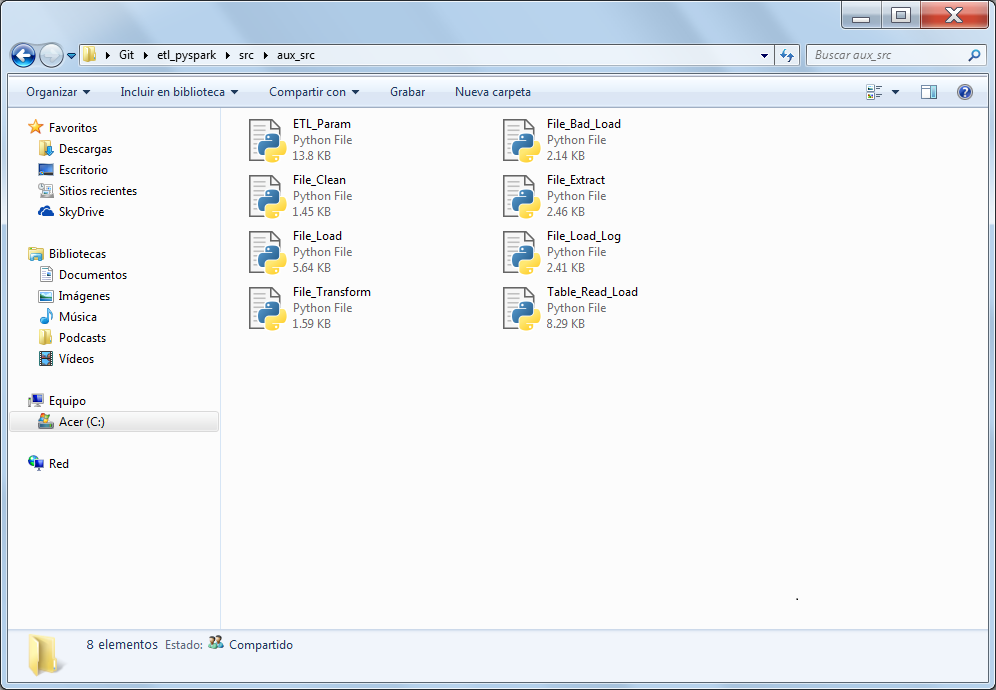
### Src directory

The src directory contains the Python files that run the Extract, Transform, and Load processes. Also, a file with some runtime errors. Also, a file with some runtime errors.



### Aux\_src directory

The aux\_src directory contains the Python files that help with the execution of the Extract, Transform and Load processes.



### Extract.py

The Extract.py file executes the process that creates the uber\_data\_yyyymmdd.csv file in the output path, which is the basis for generating the other files for loading into the tables.

To achieve this, run the next functions from the File\_Extract.py file

deleteFile()

createFile()

### Transform.py

The Extract.py file executes the process that creates the next files in the output path

cat\_tarifa\_yyyymmdd.csv

cat\_tipo\_pago\_yyyymmdd.csv

cat\_proveedor\_yyyymmdd.csv

cat\_locacion\_descenso\_yyyymmdd.csv

cat\_locacion\_recogida\_yyyymmdd.csv

dim\_fecha\_yyyymmdd.csv

fact\_pago\_viaje\_yyyymmdd.csv

To achieve this, run the next functions from the File\_Transform.py file

deleteFile()

createFile()

And gets the next parameters from the ETL\_Param.py file.

file\_tarifa\_name

file\_tipo\_pago\_name

file\_proveedor\_name

file\_locacion\_descenso\_name

file\_locacion\_recogida\_name

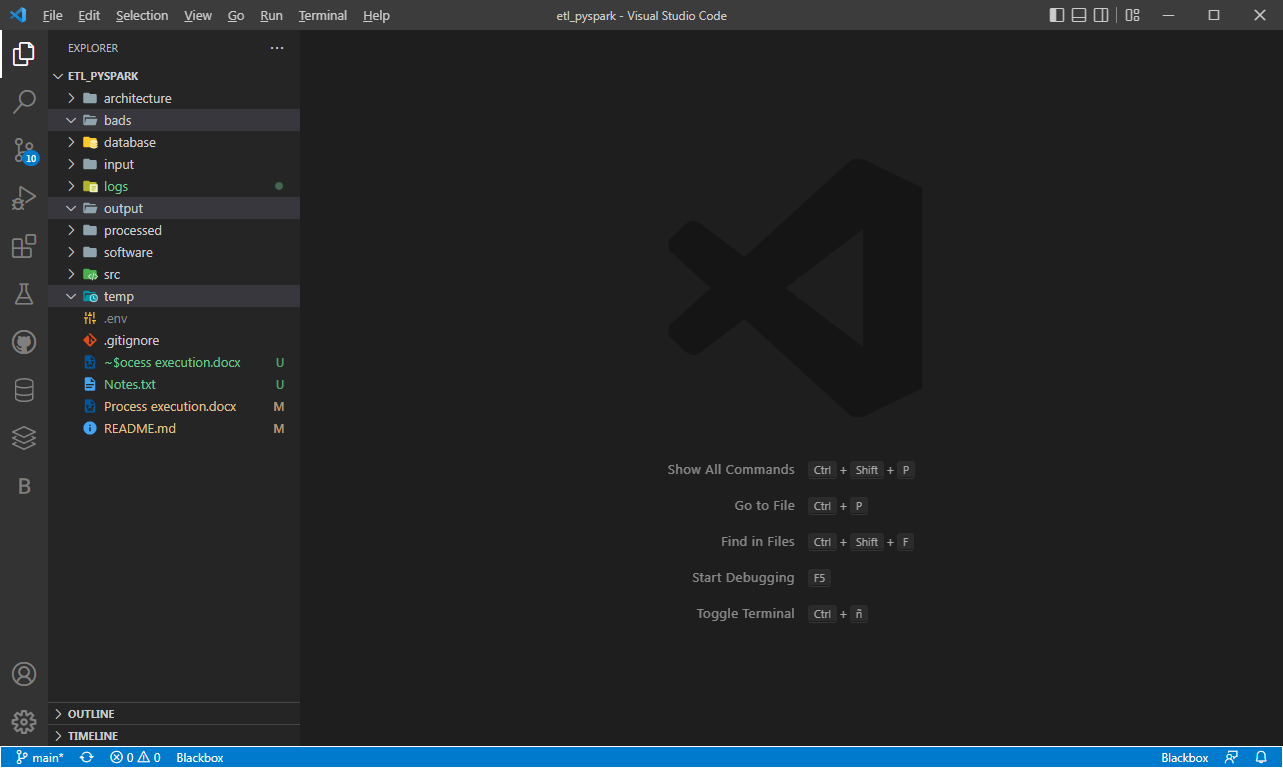
file\_fecha\_name

file\_pago\_viaje\_name

path\_transform\_log\_name

temp\_path

When the Load process finishes, the files in the bads, temp and output directories are deleted.



### Src directory

The src directory contains the python files that are executed for each of the Extract, Transform and Load processes.

