MAESTRÍA EN INTELIGENCIA ARTIFICIAL Y CIENCIA DE DATOS

CASO DE EXTRACCION TRANSFORMACION Y CARGA DE DATOS

Yoniliman Galvis Aguirre
22500214

CASO

Dataset:

Definición: Dataset de base de datos industrial en formato csv ubicado en github

Trata de la producción de 8 trenes, 51 variantes

516 días, 14876 lotes y 236530 registros

Obtener Datos (ingesta):

Se carga configuracion de archivo yaml

Se eliminan archivos dde datos mas viejos que 7 dias

Se descarga el achivo .raw desde github, se sube a un pandas df.

Se guarda en csv en la carpeta de data

Se crea un pandas profiling para analizar los datos.

Extraer datos:

Se usa archivos yaml para cargar certificados.

Se toma el csv original a un pandas df y se preprocesa para buscar los datos hurfanos, como es inferior a 2% se eliminan

Se crea la base de datos para ETL en postgresql

Se crea tabla para el dataset

Se hace un vaceado a la tabla creada

Transformar datos

Se usa archivos yaml para cargar certificados.

crea tablas de staging usando un archivo .sql en la base de datos

Se preprocesan, organizan y preprocesan los datos, se populan las tablas de stagin

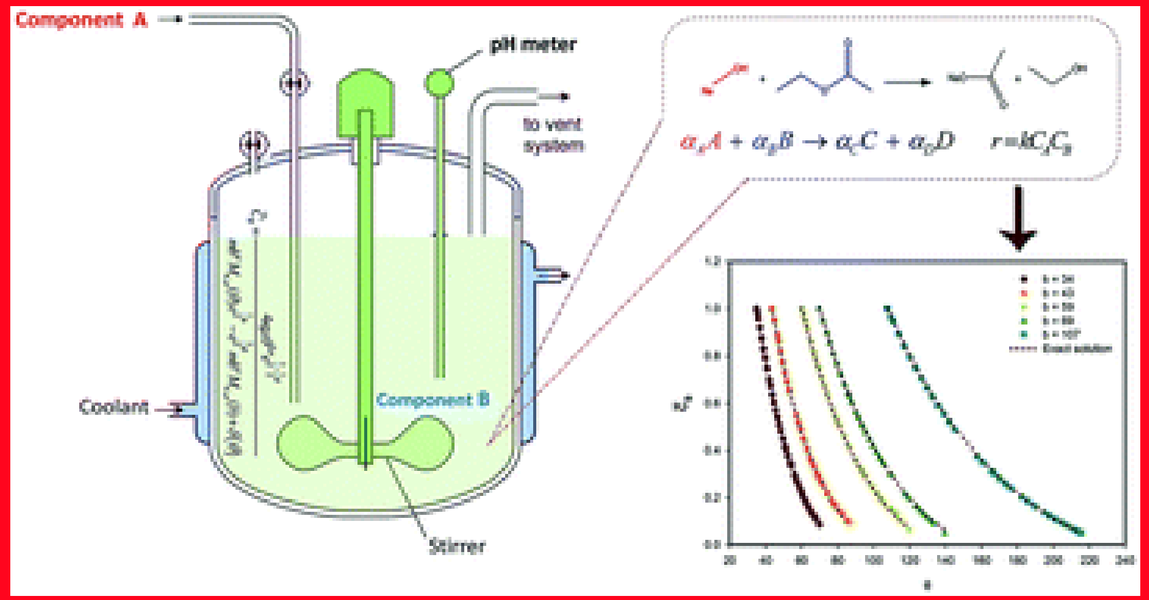
Carga

Usando METABASE se conecta a posgresql

Se leen las tablas

Se crean las preguntas (consultas)

Se crean dashboards

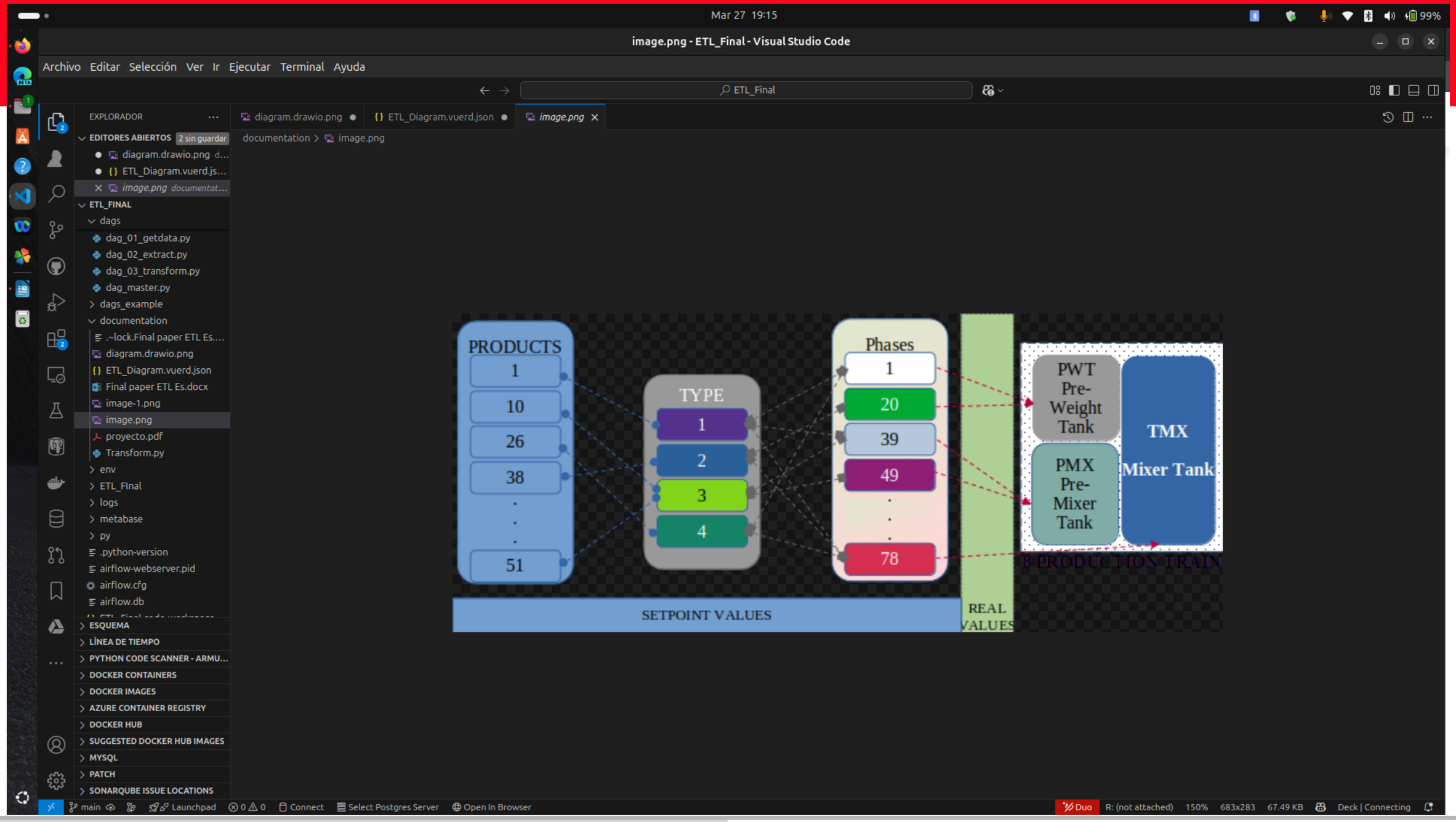


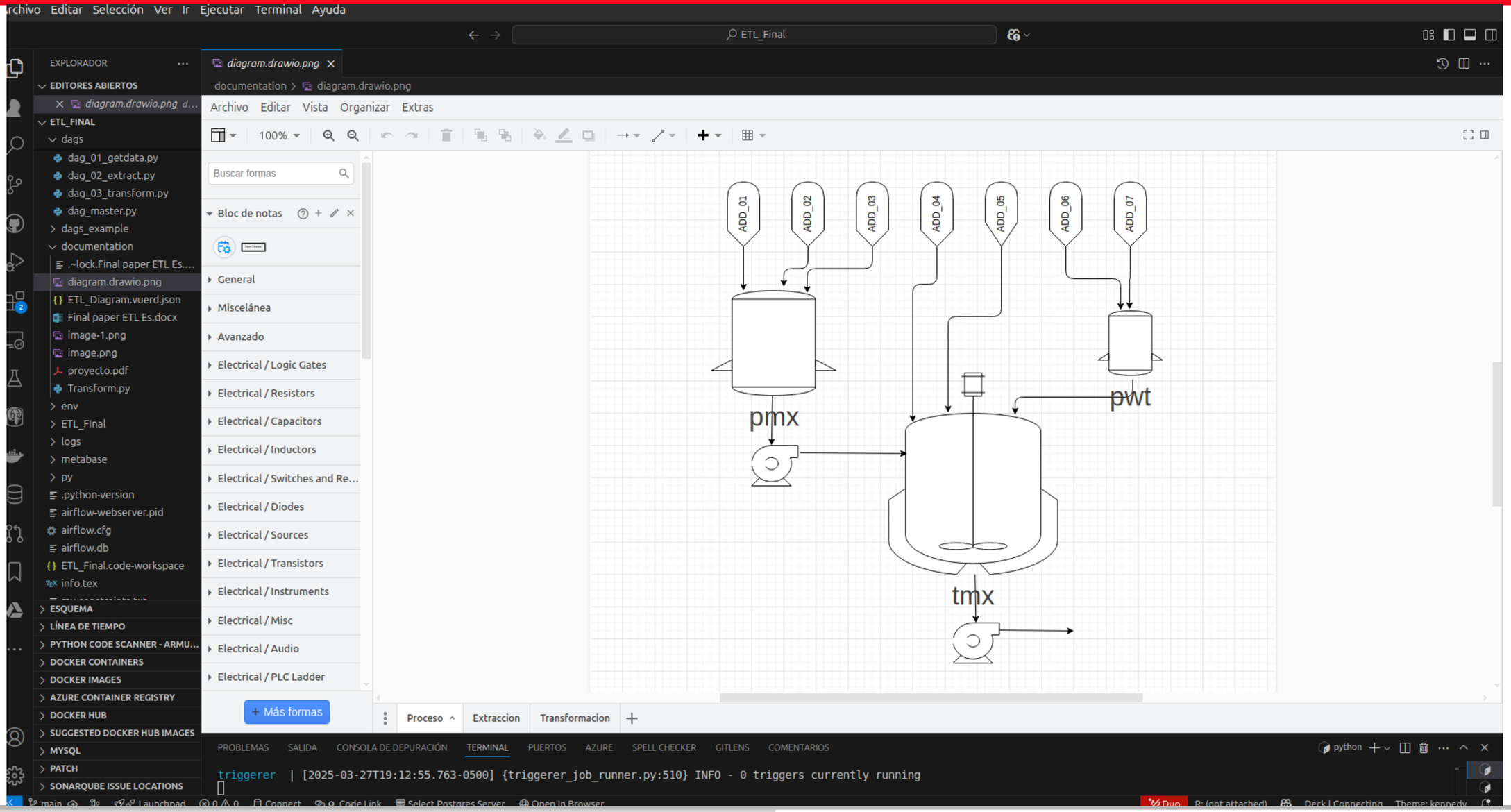
Los procesos de fabricación por lotes son sistemas con una variabilidad compleja y, especialmente en procesos de múltiples productos, garantizar la calidad del producto requiere pruebas de calidad constantes que demandan tiempo de producción, costos laborales y el uso de reactivos químicos que producen desechos que a menudo requieren un tratamiento costoso antes de que sus residuos puedan ser liberados al medio ambiente.

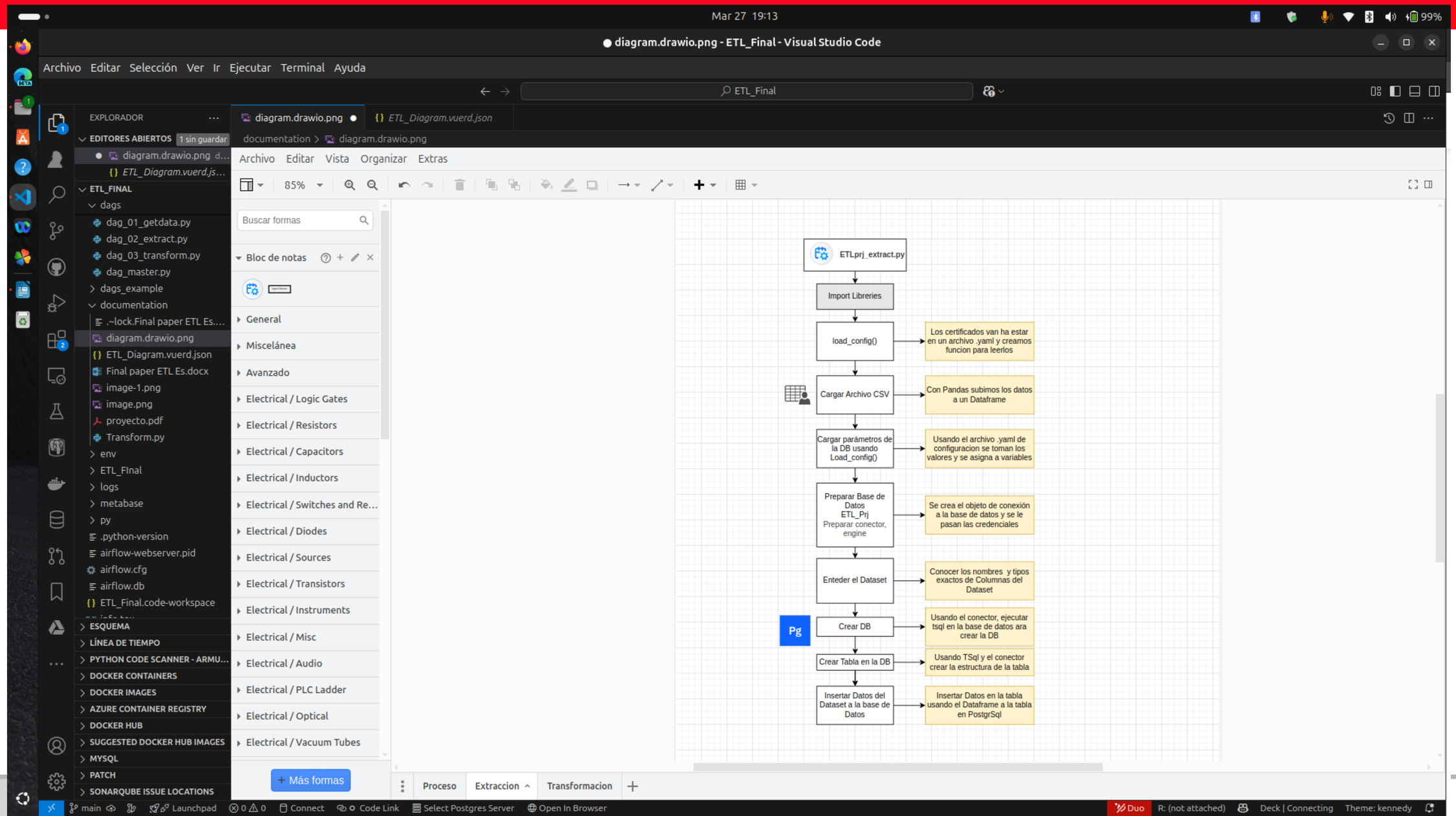
Está más que justificado utilizar herramientas de ETL, analítica de datos, inteligencia artificial y aprendizaje automático para desarrollar nuevas herramientas que mejoren los procesos de fabricación actuales con menores costos, menos horas de trabajo y un menor impacto en el medio ambiente, apuntando a un proceso sin residuos.

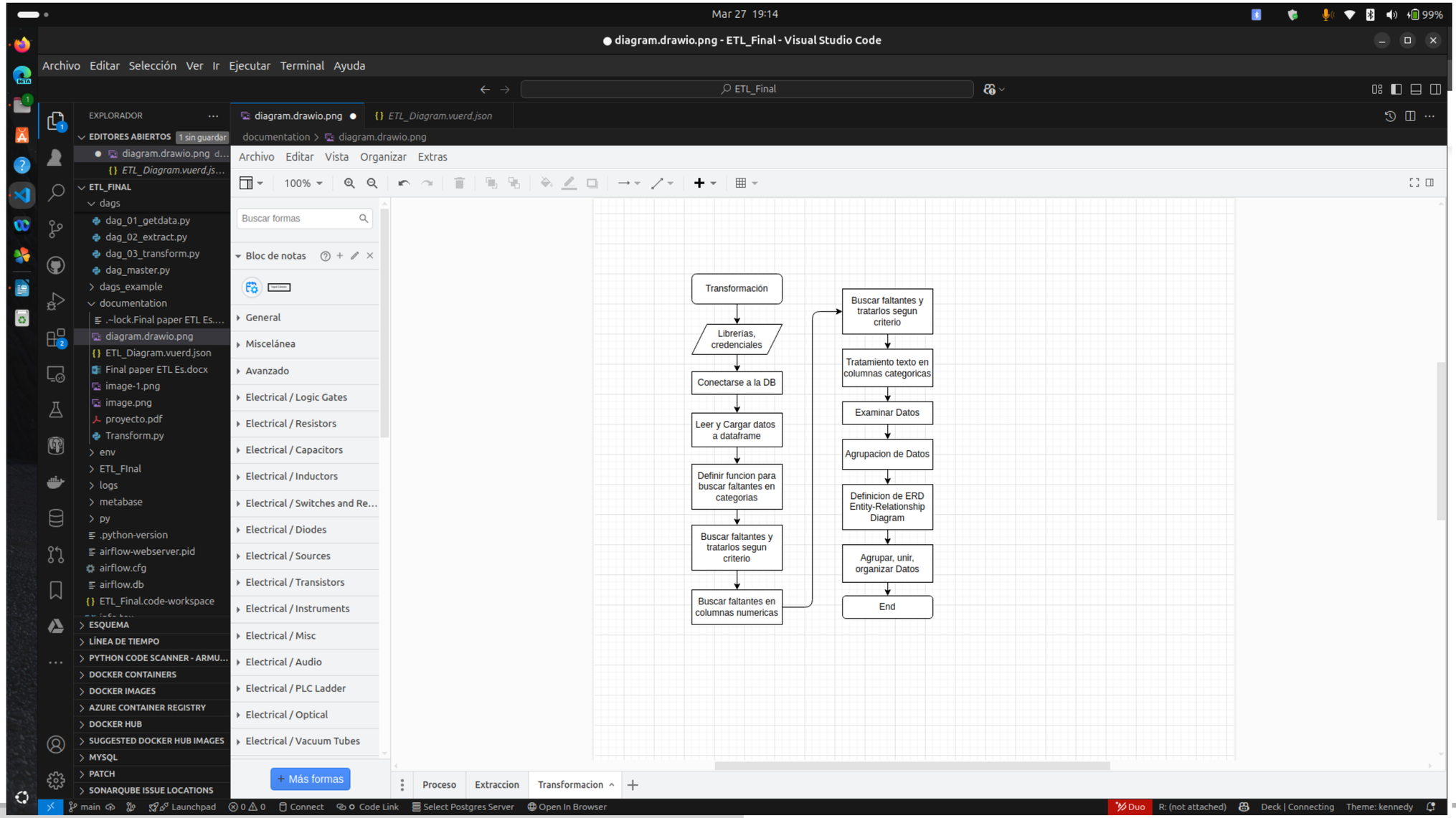
Los modelos ETL tambien permite desarrollar Dashboards que permitan analizar el comportamiento de los procesos para poder mejorarlos y reducir los fuera de especificaciones y tambien desarrollar mejores planes de produccion que disminuyan la cantidad de recursos necesarios para la fabricación.

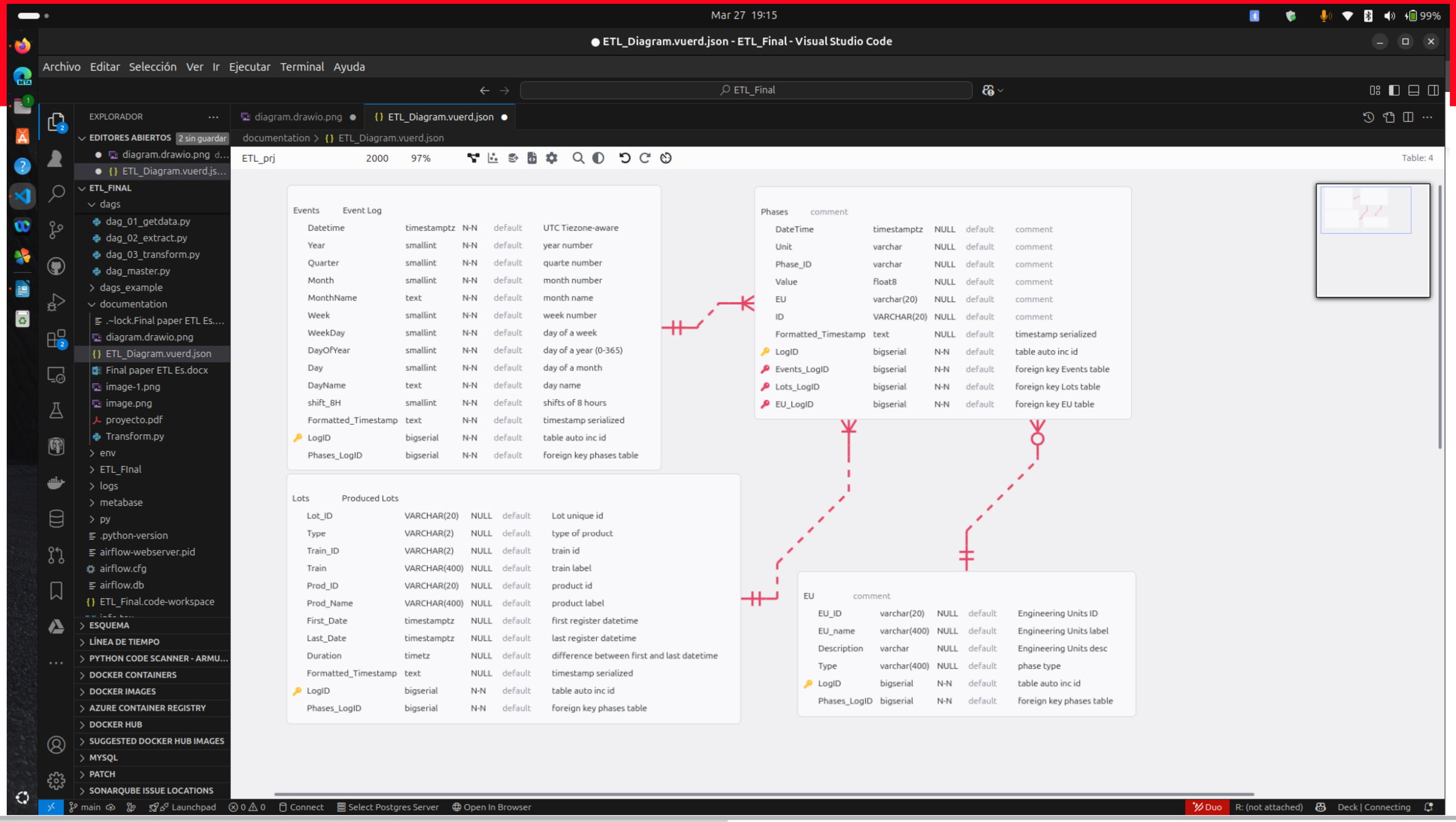
DATASET

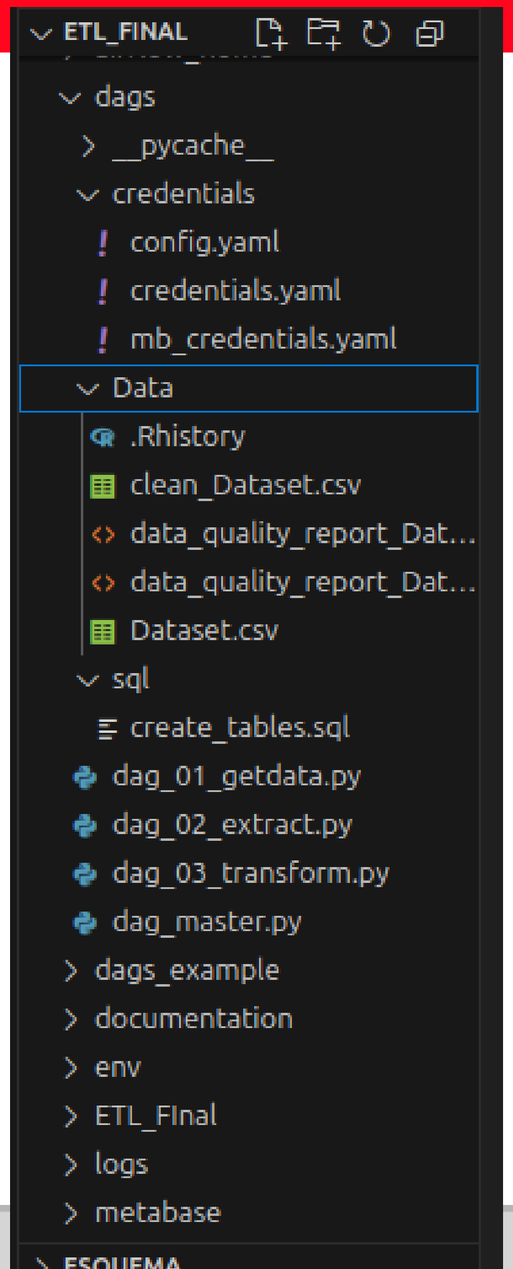

Mar 27 19:15
image.png - ETL_Final - Visual Studio Code
Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
ETL_Final
EXPLORADOR
EDITORES ABIERTOS 2 sin guardar
diagram.drawio.png
ETL_Diagram.vuerd.js...
image.png documentat...
ETL_FINAL
dags
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
.~lock.Final paper ETL Es....
diagram.drawio.png
ETL_Diagram.vuerd.json
Final paper ETL Es.docx
image-1.png
image.png
proyecto.pdf
Transform.py
env
ETL_FInal
logs
metabase
py
.python-version
airflow-webserver.pid
airflow.cfg
airflow.db
ESQUEMA
LÍNEA DE TIEMPO
PYTHON CODE SCANNER - ARMU...
DOCKER CONTAINERS
DOCKER IMAGES
AZURE CONTAINER REGISTRY
DOCKER HUB
SUGGESTED DOCKER HUB IMAGES
MYSQL
PATCH
SONARQUBE ISSUE LOCATIONS
diagram.drawio.png
ETL_Diagram.vuerd.json
image.png
documentation > image.png
PRODUCTS
1
10
26
38
51
TYPE
1
2
3
4
Phases
1
20
39
49
78
PWT Pre-Weight Tank
PMX Pre-Mixer Tank
TMX Mixer Tank
8 PRODUCTION TRAIN
SETPOINT VALUES
REAL VALUES
main
Launchpad
Connect
Select Postgres Server
Open In Browser
Duo
R: (not attached)
150%
683x283
67.49 KB
Deck | Connecting

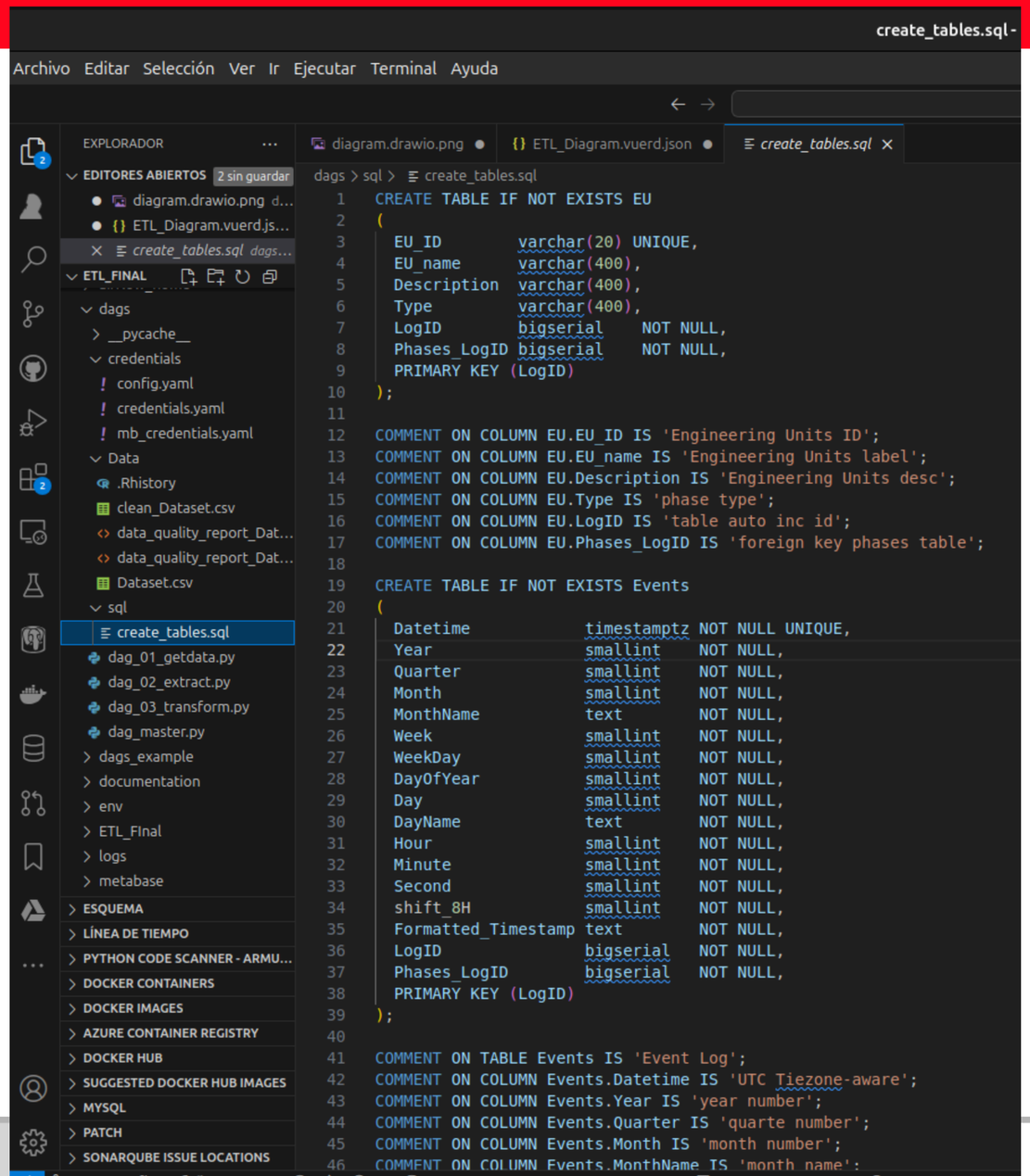

ADD_01
ADD_02
ADD_03
ADD_04
ADD_05
ADD_06
ADD_07
pmx
pwt
tmx
Proceso
Extraccion
Transformacion
triggerer | [2025-03-27T19:12:55.763-0500] {triggerer_job_runner.py:510} INFO - 0 triggers currently running

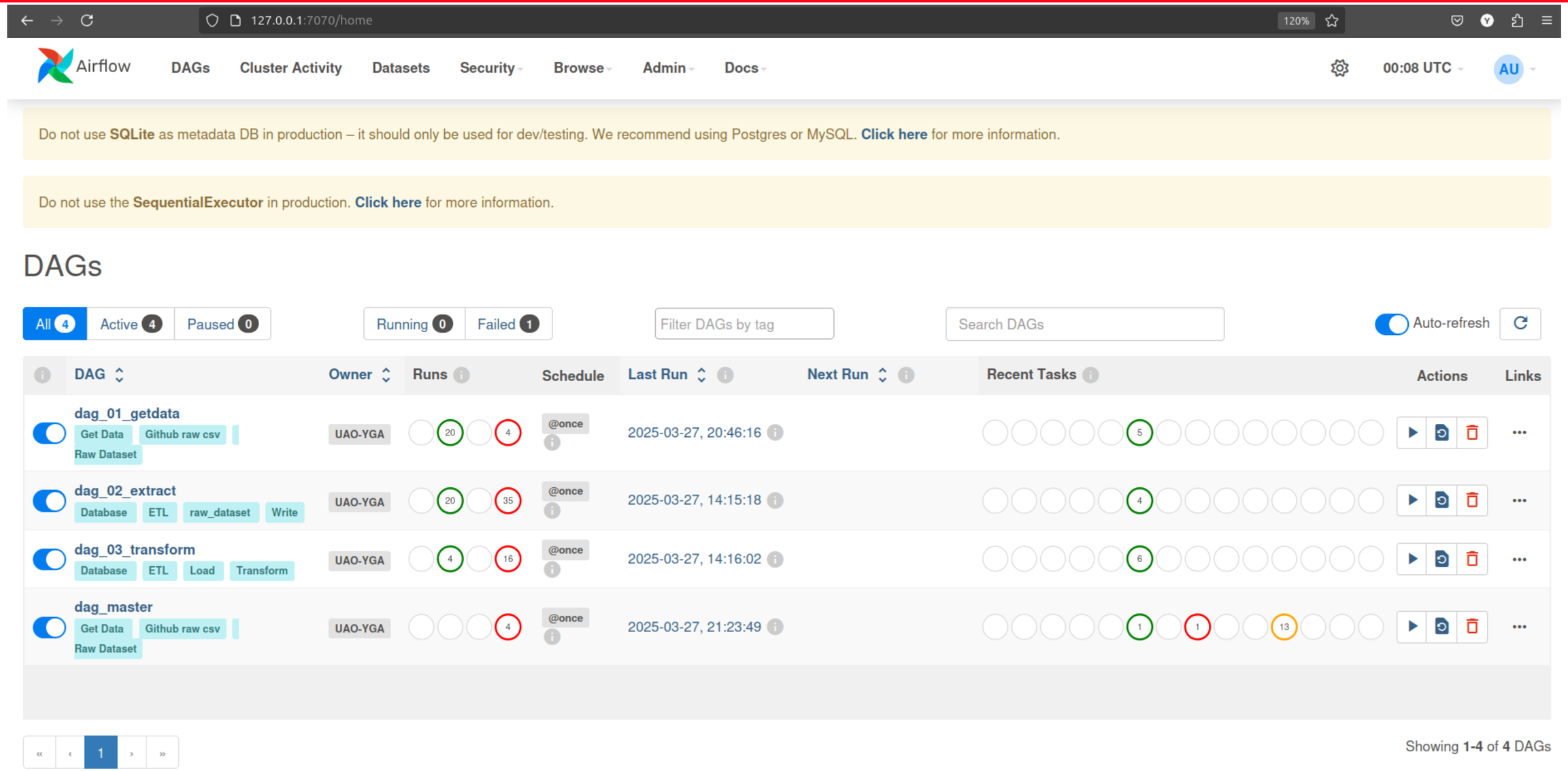

Mar 27 19:13
99%
● diagram.drawio.png - ETL_Final - Visual Studio Code
Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
ETL_Final
EXPLORADOR
EDITORES ABIERTOS 1 sin guardar
diagram.drawio.png
ETL_Diagram.vuerd.js...
ETL_FINAL
dags
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
.~lock.Final paper ETL Es....
diagram.drawio.png
ETL_Diagram.vuerd.json
Final paper ETL Es.docx
image-1.png
image.png
proyecto.pdf
Transform.py
env
ETL_Final
logs
metabase
py
.python-version
airflow-webserver.pid
airflow.cfg
airflow.db
ETL_Final.code-workspace
ESQUEMA
LÍNEA DE TIEMPO
PYTHON CODE SCANNER - ARMU...
DOCKER CONTAINERS
DOCKER IMAGES
AZURE CONTAINER REGISTRY
DOCKER HUB
SUGGESTED DOCKER HUB IMAGES
MYSQL
PATCH
SONARQUBE ISSUE LOCATIONS
diagram.drawio.png
ETL_Diagram.vuerd.json
documentation > diagram.drawio.png
Archivo Editar Vista Organizar Extras
85%
Buscar formas
Bloc de notas
General
Miscelánea
Avanzado
Electrical / Logic Gates
Electrical / Resistors
Electrical / Capacitors
Electrical / Inductors
Electrical / Switches and Re...
Electrical / Diodes
Electrical / Sources
Electrical / Transistors
Electrical / Instruments
Electrical / Misc
Electrical / Audio
Electrical / PLC Ladder
Electrical / Optical
Electrical / Vacuum Tubes
+ Más formas
ETLprj_extract.py
Import Libreries
load_config()
Los certificados van ha estar en un archivo .yaml y creamos funcion para leerlos
Cargar Archivo CSV
Con Pandas subimos los datos a un Dataframe
Cargar parámetros de la DB usando Load_config()
Usando el archivo .yaml de configuracion se toman los valores y se asigna a variables
Preparar Base de Datos ETL_Prj
Preparar conector, engine
Se crea el objeto de conexión a la base de datos y se le pasan las credenciales
Enteder el Dataset
Conocer los nombres y tipos exactos de Columnas del Dataset
Pg
Crear DB
Usando el conector, ejecutar tsql en la base de datos ara crear la DB
Crear Tabla en la DB
Usando TSql y el conector crear la estructura de la tabla
Insertar Datos del Dataset a la base de Datos
Insertar Datos en la tabla usando el Dataframe a la tabla en PostgrSql
Proceso
Extraccion
Transformacion
main
Launchpad
0 0
Connect
Code Link
Select Postgres Server
Open In Browser
Duo
R: (not attached)
Deck | Connecting
Theme: kennedy

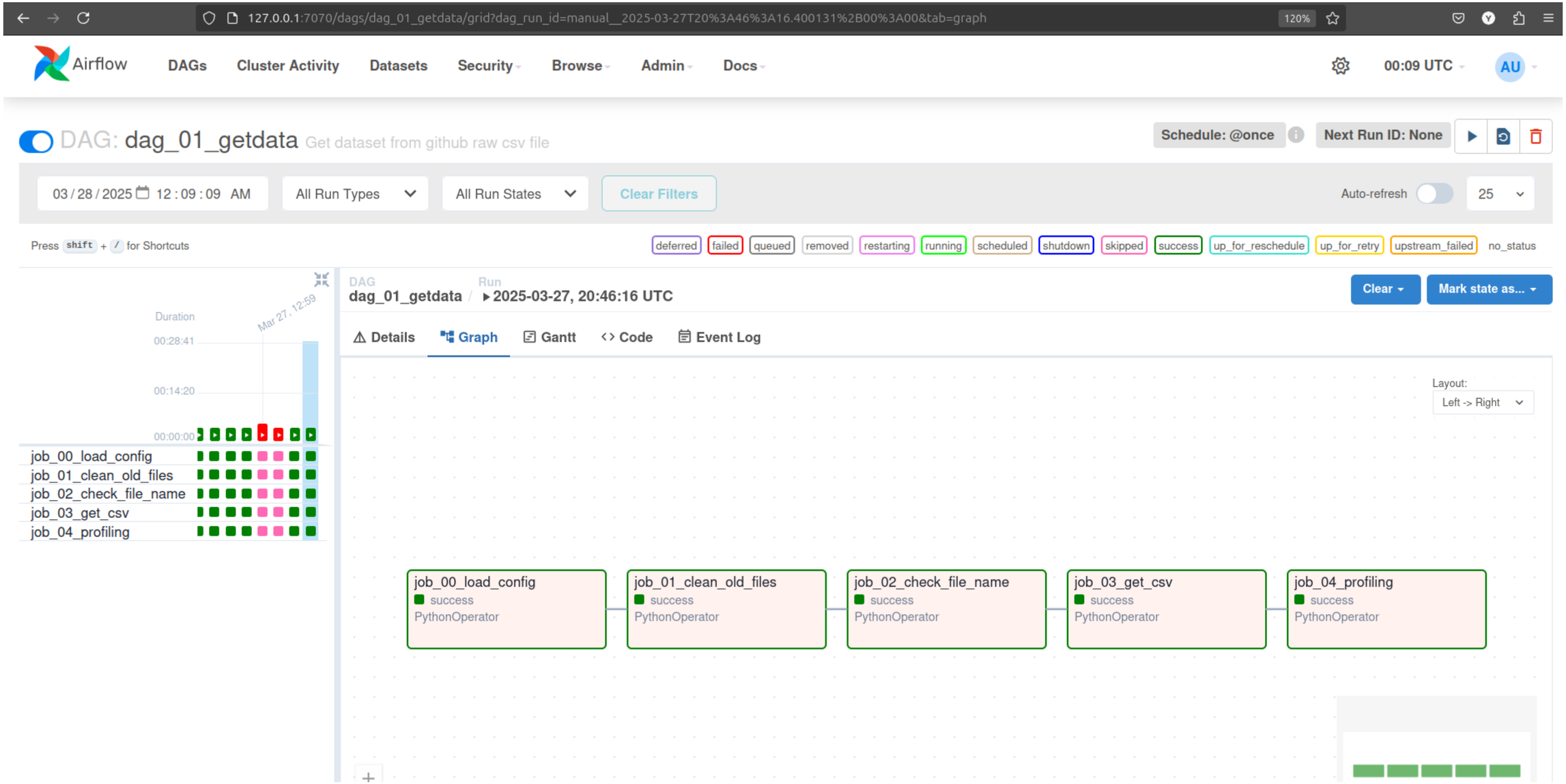

Mar 27 19:14
diagram.drawio.png - ETL_Final - Visual Studio Code
Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
ETL_Final
EXPLORADOR
EDITORES ABIERTOS 1 sin guardar
diagram.drawio.png
ETL_Diagram.vuerd.json
ETL_FINAL
dags
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
.~lock.Final paper ETL Es...
diagram.drawio.png
ETL_Diagram.vuerd.json
Final paper ETL Es.docx
image-1.png
image.png
proyecto.pdf
Transform.py
env
ETL_FInal
logs
metabase
py
.python-version
airflow-webserver.pid
airflow.cfg
airflow.db
ETL_Final.code-workspace
ESQUEMA
LÍNEA DE TIEMPO
PYTHON CODE SCANNER - ARMU...
DOCKER CONTAINERS
DOCKER IMAGES
AZURE CONTAINER REGISTRY
DOCKER HUB
SUGGESTED DOCKER HUB IMAGES
MYSQL
PATCH
SONARQUBE ISSUE LOCATIONS
documentation > diagram.drawio.png
Archivo Editar Vista Organizar Extras
100%
Buscar formas
Bloc de notas
General
Miscelánea
Avanzado
Electrical / Logic Gates
Electrical / Resistors
Electrical / Capacitors
Electrical / Inductors
Electrical / Switches and Re...
Electrical / Diodes
Electrical / Sources
Electrical / Transistors
Electrical / Instruments
Electrical / Misc
Electrical / Audio
Electrical / PLC Ladder
Electrical / Optical
Electrical / Vacuum Tubes
+ Más formas
Transformación
Librerías, credenciales
Conectarse a la DB
Leer y Cargar datos a dataframe
Definir funcion para buscar faltantes en categorias
Buscar faltantes y tratarlos segun criterio
Buscar faltantes en columnas numericas
Buscar faltantes y tratarlos segun criterio
Tratamiento texto en columnas categoricas
Examinar Datos
Agrupacion de Datos
Definicion de ERD Entity-Relationship Diagram
Agrupar, unir, organizar Datos
End
Proceso
Extraccion
Transformacion
main
Launchpad
Connect
Code Link
Select Postgres Server
Open In Browser
Duo
R: (not attached)
Deck | Connecting
Theme: kennedy

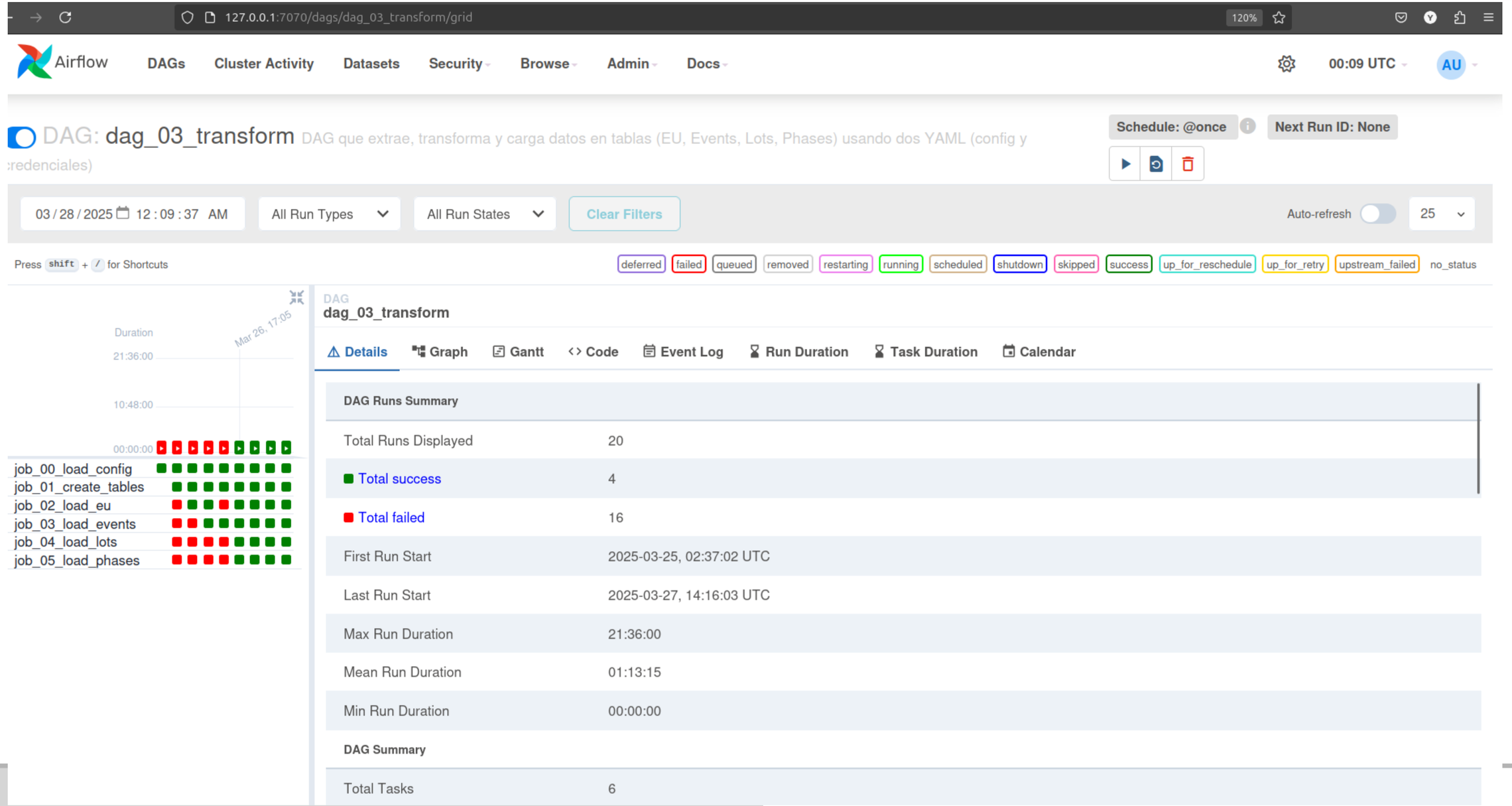

Mar 27 19:15
99%
● ETL_Diagram.vuerd.json - ETL_Final - Visual Studio Code
Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
ETL_Final
EXPLORADOR
EDITORES ABIERTOS 2 sin guardar
diagram.drawio.png
ETL_Diagram.vuerd.js...
ETL_FINAL
dags
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
.~lock.Final paper ETL Es....
diagram.drawio.png
ETL_Diagram.vuerd.json
Final paper ETL Es.docx
image-1.png
image.png
proyecto.pdf
Transform.py
env
ETL_FInal
logs
metabase
py
.python-version
airflow-webserver.pid
airflow.cfg
airflow.db
ETL_Final.code-workspace
ESQUEMA
LÍNEA DE TIEMPO
PYTHON CODE SCANNER - ARMU...
DOCKER CONTAINERS
DOCKER IMAGES
AZURE CONTAINER REGISTRY
DOCKER HUB
SUGGESTED DOCKER HUB IMAGES
MYSQL
PATCH
SONARQUBE ISSUE LOCATIONS
diagram.drawio.png
ETL_Diagram.vuerd.json
documentation > ETL_Diagram.vuerd.json
ETL_prj 2000 97%
Table: 4
Events Event Log
Datetime timestamptz N-N default UTC Tiezone-aware
Year smallint N-N default year number
Quarter smallint N-N default quarte number
Month smallint N-N default month number
MonthName text N-N default month name
Week smallint N-N default week number
WeekDay smallint N-N default day of a week
DayOfYear smallint N-N default day of a year (0-365)
Day smallint N-N default day of a month
DayName text N-N default day name
shift_8H smallint N-N default shifts of 8 hours
Formatted_Timestamp text N-N default timestamp serialized
LogID bigserial N-N default table auto inc id
Phases_LogID bigserial N-N default foreign key phases table
Phases comment
DateTime timestamptz NULL default comment
Unit varchar NULL default comment
Phase_ID varchar NULL default comment
Value float8 NULL default comment
EU varchar(20) NULL default comment
ID VARCHAR(20) NULL default comment
Formatted_Timestamp text NULL default timestamp serialized
LogID bigserial N-N default table auto inc id
Events_LogID bigserial N-N default foreign key Events table
Lots_LogID bigserial N-N default foreign key Lots table
EU_LogID bigserial N-N default foreign key EU table
Lots Produced Lots
Lot_ID VARCHAR(20) NULL default Lot unique id
Type VARCHAR(2) NULL default type of product
Train_ID VARCHAR(2) NULL default train id
Train VARCHAR(400) NULL default train label
Prod_ID VARCHAR(20) NULL default product id
Prod_Name VARCHAR(400) NULL default product label
First_Date timestamptz NULL default first register datetime
Last_Date timestamptz NULL default last register datetime
Duration timetz NULL default difference between first and last datetime
Formatted_Timestamp text NULL default timestamp serialized
LogID bigserial N-N default table auto inc id
Phases_LogID bigserial N-N default foreign key phases table
EU comment
EU_ID varchar(20) NULL default Engineering Units ID
EU_name varchar(400) NULL default Engineering Units label
Description varchar NULL default Engineering Units desc
Type varchar(400) NULL default phase type
LogID bigserial N-N default table auto inc id
Phases_LogID bigserial N-N default foreign key phases table
main
Launchpad
0 0
Connect
Select Postgres Server
Open In Browser
Duo
R: (not attached)
Deck | Connecting

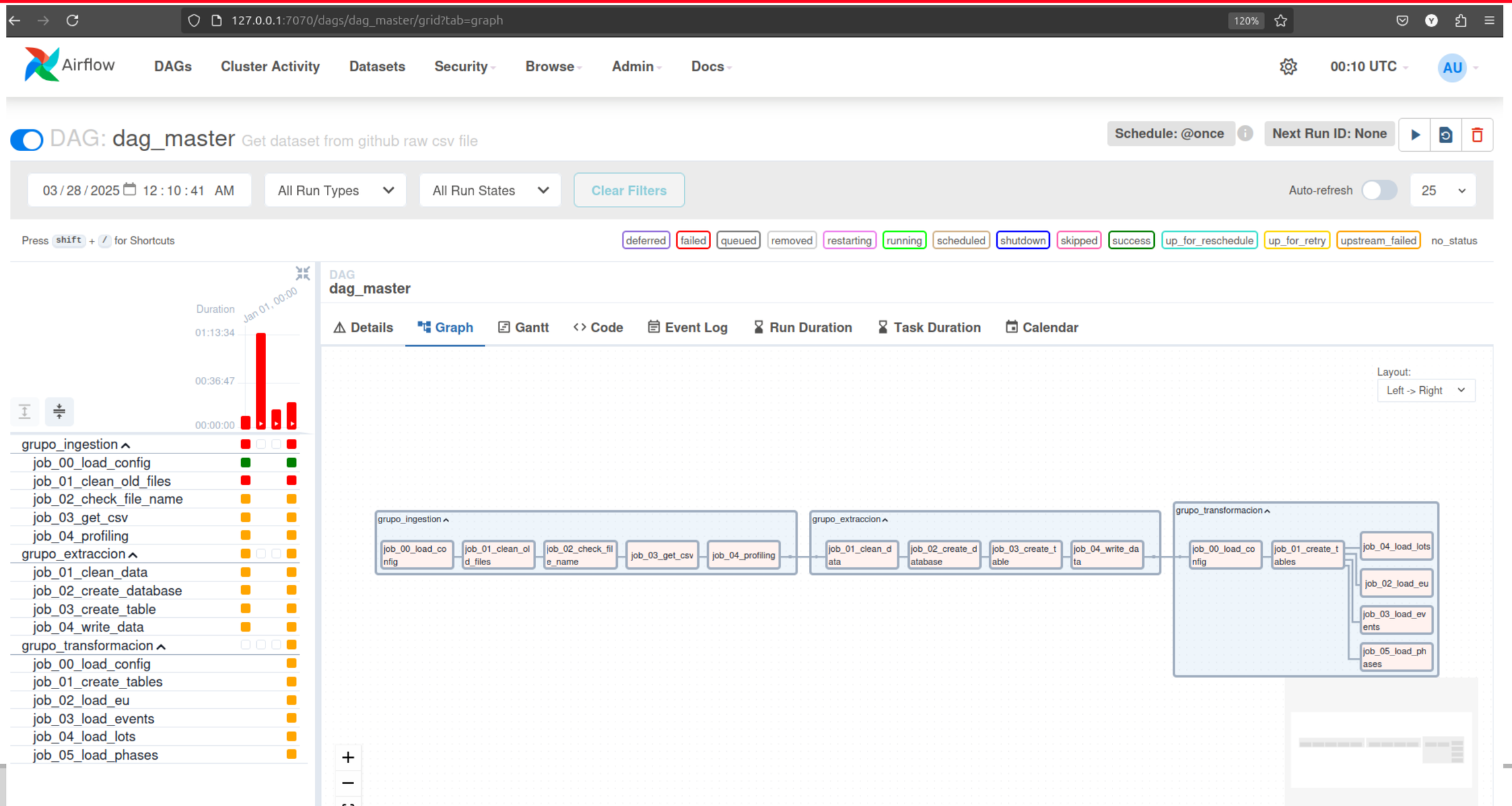

ETL_FINAL
dags
__pycache__
credentials
config.yaml
credentials.yaml
mb_credentials.yaml
Data
.Rhistory
clean_Dataset.csv
data_quality_report_Dat...
data_quality_report_Dat...
Dataset.csv
sql
create_tables.sql
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
env
ETL_FInal
logs
metabase
ESQUEMA

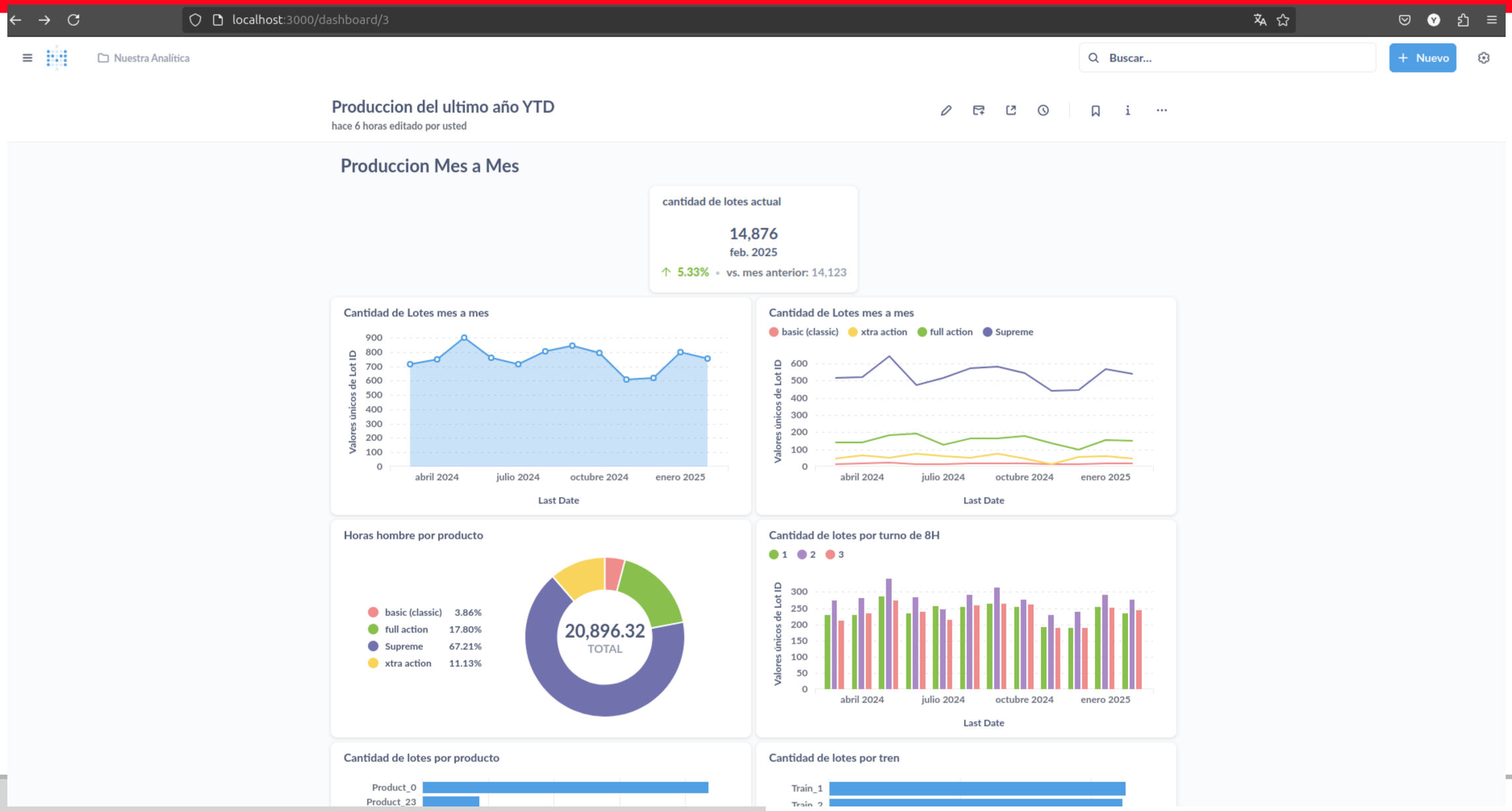

create_tables.sql -
Archivo Editar Selección Ver Ir Ejecutar Terminal Ayuda
EXPLORADOR
EDITORES ABIERTOS 2 sin guardar
diagram.drawio.png
ETL_Diagram.vuerd.js...
create_tables.sql dags...
ETL_FINAL
dags
__pycache__
credentials
config.yaml
credentials.yaml
mb_credentials.yaml
Data
.Rhistory
clean_Dataset.csv
data_quality_report_Dat...
data_quality_report_Dat...
Dataset.csv
sql
create_tables.sql
dag_01_getdata.py
dag_02_extract.py
dag_03_transform.py
dag_master.py
dags_example
documentation
env
ETL_FInal
logs
metabase
ESQUEMA
LÍNEA DE TIEMPO
PYTHON CODE SCANNER - ARMU...
DOCKER CONTAINERS
DOCKER IMAGES
AZURE CONTAINER REGISTRY
DOCKER HUB
SUGGESTED DOCKER HUB IMAGES
MYSQL
PATCH
SONARQUBE ISSUE LOCATIONS
diagram.drawio.png
ETL_Diagram.vuerd.json
create_tables.sql
dags > sql > create_tables.sql
1 CREATE TABLE IF NOT EXISTS EU
2 (
3 EU_ID varchar(20) UNIQUE,
4 EU_name varchar(400),
5 Description varchar(400),
6 Type varchar(400),
7 LogID bigserial NOT NULL,
8 Phases_LogID bigserial NOT NULL,
9 PRIMARY KEY (LogID)
10);
11
12 COMMENT ON COLUMN EU.EU_ID IS 'Engineering Units ID';
13 COMMENT ON COLUMN EU.EU_name IS 'Engineering Units label';
14 COMMENT ON COLUMN EU.Description IS 'Engineering Units desc';
15 COMMENT ON COLUMN EU.Type IS 'phase type';
16 COMMENT ON COLUMN EU.LogID IS 'table auto inc id';
17 COMMENT ON COLUMN EU.Phases_LogID IS 'foreign key phases table';
18
19 CREATE TABLE IF NOT EXISTS Events
20 (
21 Datetime timestamptz NOT NULL UNIQUE,
22 Year smallint NOT NULL,
23 Quarter smallint NOT NULL,
24 Month smallint NOT NULL,
25 MonthName text NOT NULL,
26 Week smallint NOT NULL,
27 WeekDay smallint NOT NULL,
28 DayOfYear smallint NOT NULL,
29 Day smallint NOT NULL,
30 DayName text NOT NULL,
31 Hour smallint NOT NULL,
32 Minute smallint NOT NULL,
33 Second smallint NOT NULL,
34 shift_8H smallint NOT NULL,
35 Formatted_Timestamp text NOT NULL,
36 LogID bigserial NOT NULL,
37 Phases_LogID bigserial NOT NULL,
38 PRIMARY KEY (LogID)
39);
40
41 COMMENT ON TABLE Events IS 'Event Log';
42 COMMENT ON COLUMN Events.Datetime IS 'UTC Tiezone-aware';
43 COMMENT ON COLUMN Events.Year IS 'year number';
44 COMMENT ON COLUMN Events.Quarter IS 'quarte number';
45 COMMENT ON COLUMN Events.Month IS 'month number';
46 COMMENT ON COLUMN Events.MonthName IS 'month name';


127.0.0.1:7070/home
120%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
00:08 UTC
AU
Do not use SQLite as metadata DB in production – it should only be used for dev/testing. We recommend using Postgres or MySQL. Click here for more information.
Do not use the SequentialExecutor in production. Click here for more information.
DAGs
All 4
Active 4
Paused 0
Running 0
Failed 1
Filter DAGs by tag
Search DAGs
Auto-refresh
DAG
Owner
Runs
Schedule
Last Run
Next Run
Recent Tasks
Actions
Links
dag_01_getdata
Get Data
Github raw csv
Raw Dataset
UAO-YGA
@once
2025-03-27, 20:46:16
dag_02_extract
Database
ETL
raw_dataset
Write
UAO-YGA
@once
2025-03-27, 14:15:18
dag_03_transform
Database
ETL
Load
Transform
UAO-YGA
@once
2025-03-27, 14:16:02
dag_master
Get Data
Github raw csv
Raw Dataset
UAO-YGA
@once
2025-03-27, 21:23:49
1
Showing 1-4 of 4 DAGs


127.0.0.1:7070/dags/dag_01_getdata/grid?dag_run_id=manual__2025-03-27T20%3A46%3A16.400131%2B00%3A00&tab=graph
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
00:09 UTC
AU
DAG: dag_01_getdata
Get dataset from github raw csv file
Schedule: @once
Next Run ID: None
03 / 28 / 2025 12 : 09 : 09 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Mar 27, 12:59
00:28:41
00:14:20
00:00:00
job_00_load_config
job_01_clean_old_files
job_02_check_file_name
job_03_get_csv
job_04_profiling
DAG
dag_01_getdata
Run
2025-03-27, 20:46:16 UTC
Clear
Mark state as...
Details
Graph
Gantt
Code
Event Log
Layout:
Left -> Right
job_00_load_config
success
PythonOperator
job_01_clean_old_files
success
PythonOperator
job_02_check_file_name
success
PythonOperator
job_03_get_csv
success
PythonOperator
job_04_profiling
success
PythonOperator


127.0.0.1:7070/dags/dag_03_transform/grid
120%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
00:09 UTC
AU
DAG: dag_03_transform
DAG que extrae, transforma y carga datos en tablas (EU, Events, Lots, Phases) usando dos YAML (config y
credenciales)
Schedule: @once
Next Run ID: None
03 / 28 / 2025 12 : 09 : 37 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Mar 26, 17:05
21:36:00
10:48:00
00:00:00
job_00_load_config
job_01_create_tables
job_02_load_eu
job_03_load_events
job_04_load_lots
job_05_load_phases
DAG
dag_03_transform
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
DAG Runs Summary
Total Runs Displayed 20
Total success 4
Total failed 16
First Run Start 2025-03-25, 02:37:02 UTC
Last Run Start 2025-03-27, 14:16:03 UTC
Max Run Duration 21:36:00
Mean Run Duration 01:13:15
Min Run Duration 00:00:00
DAG Summary
Total Tasks 6


127.0.0.1:7070/dags/dag_master/grid?tab=graph
120%
Airflow
DAGs
Cluster Activity
Datasets
Security
Browse
Admin
Docs
00:10 UTC
AU
DAG: dag_master
Get dataset from github raw csv file
Schedule: @once
Next Run ID: None
03 / 28 / 2025 12 : 10 : 41 AM
All Run Types
All Run States
Clear Filters
Auto-refresh
25
Press shift + / for Shortcuts
deferred
failed
queued
removed
restarting
running
scheduled
shutdown
skipped
success
up_for_reschedule
up_for_retry
upstream_failed
no_status
Duration
Jan 01, 00:00
01:13:34
00:36:47
00:00:00
grupo_ingestion
job_00_load_config
job_01_clean_old_files
job_02_check_file_name
job_03_get_csv
job_04_profiling
grupo_extraccion
job_01_clean_data
job_02_create_database
job_03_create_table
job_04_write_data
grupo_transformacion
job_00_load_config
job_01_create_tables
job_02_load_eu
job_03_load_events
job_04_load_lots
job_05_load_phases
DAG
dag_master
Details
Graph
Gantt
Code
Event Log
Run Duration
Task Duration
Calendar
Layout:
Left -> Right
grupo_ingestion
job_00_load_config
job_01_clean_old_files
job_02_check_file_name
job_03_get_csv
job_04_profiling
grupo_extraccion
job_01_clean_data
job_02_create_database
job_03_create_table
job_04_write_data
grupo_transformacion
job_00_load_config
job_01_create_tables
job_04_load_lots
job_02_load_eu
job_03_load_events
job_05_load_phases


localhost:3000/dashboard/3
Nuestra Analítica
Buscar...
Nuevo
Produccion del ultimo año YTD
hace 6 horas editado por usted
Produccion Mes a Mes
cantidad de lotes actual
14,876
feb. 2025
5.33% • vs. mes anterior: 14,123
Cantidad de Lotes mes a mes
Valores únicos de Lot ID
900
800
700
600
500
400
300
200
100
0
abril 2024
julio 2024
octubre 2024
enero 2025
Last Date
Cantidad de Lotes mes a mes
basic (classic)
xtra action
full action
Supreme
600
500
400
300
200
100
0
abril 2024
julio 2024
octubre 2024
enero 2025
Last Date
Horas hombre por producto
basic (classic) 3.86%
full action 17.80%
Supreme 67.21%
xtra action 11.13%
20,896.32
TOTAL
Cantidad de lotes por turno de 8H
1
2
3
300
250
200
150
100
50
0
abril 2024
julio 2024
octubre 2024
enero 2025
Last Date
Cantidad de lotes por producto
Product_0
Product_23
Cantidad de lotes por tren
Train_1
Train_2


uao