# WORKSHOP # 1 - ETL

### By: Carol Varela

## 1. Introduction

This workshop aims to migrate and transform candidate data from a CSV file to a PostgreSQL database, followed by exploratory analysis and data visualization. The goal is to assess knowledge in data management and visualizations through the creation of specific metrics in charts.

## 2. Environment Setup

**Requirements**

- Python (recommended version: 3.8 or higher)
- PostgreSQL (installed and running, with the database created for this workshop)
- Python packages (installed via **requirements.txt**)

### Instalación

1. Clona el repositorio del workshop.

2. Crea un entorno virtual:

```
python -m venv venv
.\venv\Scripts\Activate.ps1
```

3. Instala los paquetes requeridos:

```
pip install -r requirements.txt
```

4. Configura las variables de entorno en un archivo `.env` con los siguientes detalles:
makefile

```
PGDIALECT=your_host
PGUSER=your_username
PGPASSWD=your_password
PGHOST=your_host_adress
PGPORT=5432
PGDB=your_db_name
WORK_DIR=/path/to/your/project
```

# 3. Workshop Structure

**src/**: Contains Python modules for database connection, ORM models, and data transformation.

- **db_connection.py:** Configuration and setup of the PostgreSQL database connection.
- **model.py:** Definition of SQLAlchemy models for the candidates table and transformed data.
- **transform.py:** Data transformation and preparation from a CSV file.

**notebooks/**: Jupyter notebooks for data migration, transformation, and exploratory analysis.

- **Data_migration.ipynb:** Migration of data from CSV to PostgreSQL database.
- **Data_transformation.ipynb:** Data transformation and loading into the transformed data table.
- **EDA.ipynb:** Exploratory data analysis of the data loaded into the database.

# 4. Modules

**db_connection.py** This module constructs an SQLAlchemy engine to connect to the PostgreSQL database using environment variable configurations.

**Functions**

- **build_engine():** Creates and returns an SQLAlchemy engine for the database.

**model.py** Defines SQLAlchemy models for the Candidates and Candidates_transformed tables.

**Classes**

- **Candidates:** Model to store candidate data.
- **Candidates_transformed:** Model to store transformed data with an additional 'Hired' column.

**transform.py** Provides the Transform class for cleaning and transforming data read from a CSV file.

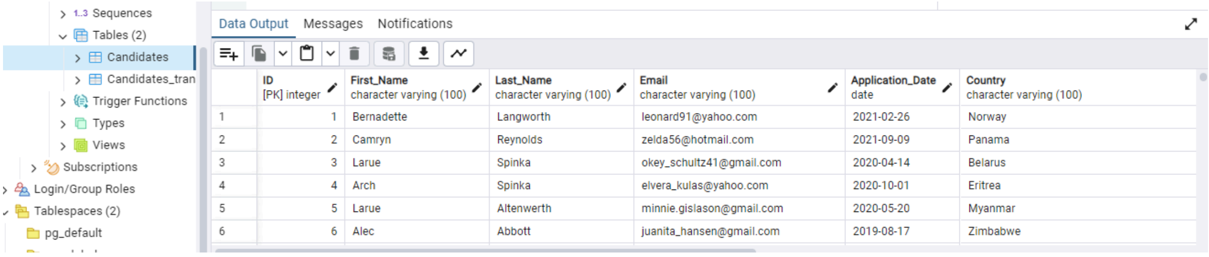
**Methods**

- **__init__(self, file):** Loads the DataFrame from a CSV file.
- **rename_columns(self):** Renames columns in the DataFrame.
- **insert_ids(self):** Adds an 'ID' column to the DataFrame.
- **add_hired_column(self):** Adds a 'Hired' column based on scores.
- **technology_to_category(self):** Maps technologies to categories.

# 5. Notebooks

**Data_migration.ipynb** This notebook performs the migration of data from a CSV file to a PostgreSQL database. Includes:

- Environment setup and module loading.
- Database connection and creation of the **Candidates table.**
- Some data transformations and loading into the database.



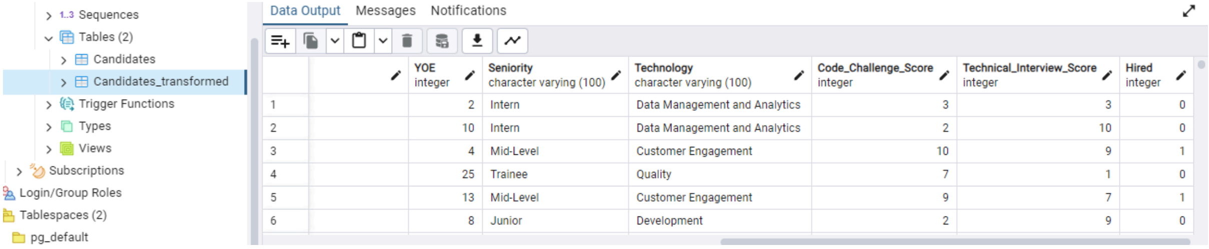
| | ID<br>[PK] integer | First_Name<br>character varying (100) | Last_Name<br>character varying (100) | Email<br>character varying (100) | Application_Date<br>date | Country<br>character varying (100) |
|---|---|---|---|---|---|---|
| 1 | 1 | Bernadette | Langworth | leonard91@yahoo.com | 2021-02-26 | Norway |
| 2 | 2 | Camryn | Reynolds | zelda56@hotmail.com | 2021-09-09 | Panama |
| 3 | 3 | Larue | Spinka | okey_schultz41@gmail.com | 2020-04-14 | Belarus |
| 4 | 4 | Arch | Spinka | elvera_kulas@yahoo.com | 2020-10-01 | Eritrea |
| 5 | 5 | Larue | Altenwerth | minnie.gislason@gmail.com | 2020-05-20 | Myanmar |
| 6 | 6 | Alec | Abbott | juanita_hansen@gmail.com | 2019-08-17 | Zimbabwe |

**EDA.ipynb** Performs exploratory data analysis on the data loaded into the database. Includes:

- Data loading and visualization.
- Identification of missing data and data types.
- Data distribution and correlations.
- Visualizations and data quality evaluation.

**Data_transformation.ipynb** Transforms data and loads it into the **Candidates_transformed** table. Includes:

- Environment setup and module loading.
- Database connection and creation of the **Candidates_transformed** table.
- Application of data transformations.



| | | YOE<br>integer | Seniority<br>character varying (100) | Technology<br>character varying (100) | Code_Challenge_Score<br>integer | Technical_Interview_Score<br>integer | Hired<br>integer |
|---|---|---|---|---|---|---|---|
| 1 | | 2 | Intern | Data Management and Analytics | 3 | 3 | 0 |
| 2 | | 10 | Intern | Data Management and Analytics | 2 | 10 | 0 |
| 3 | | 4 | Mid-Level | Customer Engagement | 10 | 9 | 1 |
| 4 | | 25 | Trainee | Quality | 7 | 1 | 0 |
| 5 | | 13 | Mid-Level | Customer Engagement | 9 | 7 | 1 |
| 6 | | 8 | Junior | Development | 2 | 9 | 0 |

# 6. Execution Procedures

- **Data Migration**: Run **Data_migration.ipynb** to load data from the CSV into the **Candidates** table.
- **Exploratory Analysis:** Run **EDA.ipynb** to explore and analyze the data.
- **Data Transformation:** Run **Data_transformation.ipynb** to transform and load data into the **Candidates_transformed** table.

# 7. Creating the Dashboard in Power BI

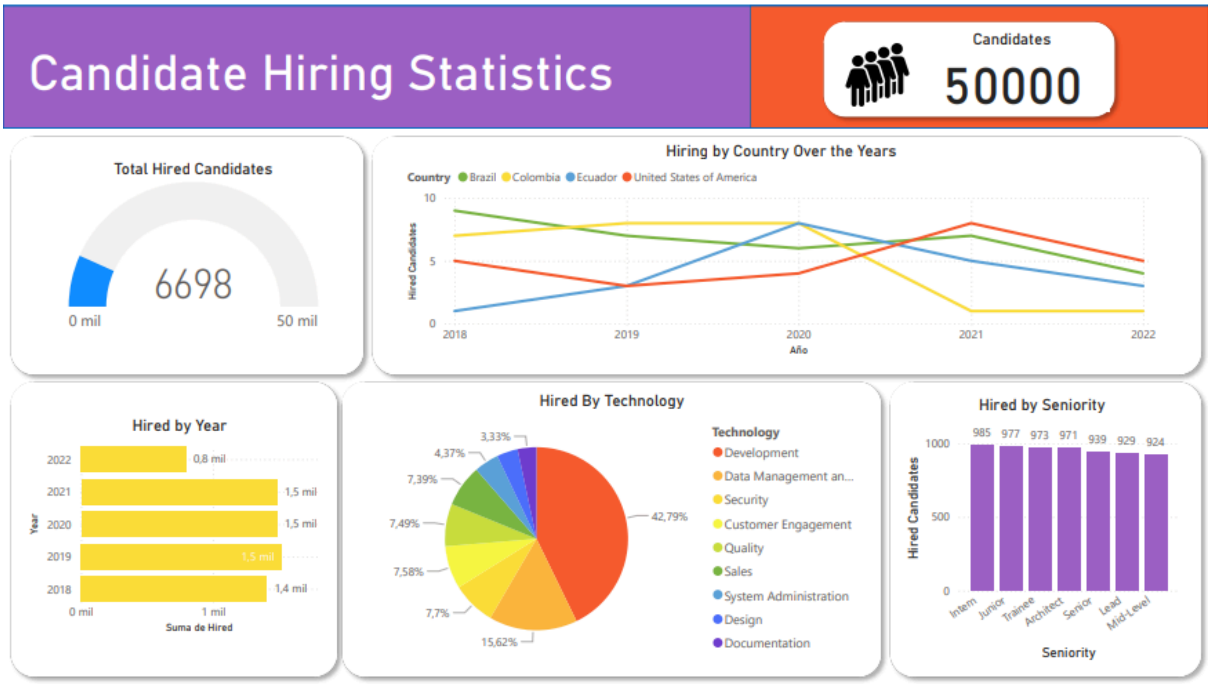
This section details the process of creating an interactive dashboard using Power BI. The dashboard is designed to provide a comprehensive visualization of key metrics extracted from candidate data.

**7.1 Dashboard Objectives** The Power BI dashboard was created to visualize the following key metrics:

- **Hires by Technology:** Distribution of hires according to the technologies used.
- **Hires by Year:** Number of hires made each year.
- **Hires by Experience Level:** Classification of hires by candidate experience level.
- **Hires by Country Over Time:** Evolution of the number of hires in selected countries over time.

**7.2 Creation Process**

1. **Data Preparation** Data was exported from the PostgreSQL database to Power BI using the database connector.
2. **Data Import into Power BI** The PostgreSQL connector in Power BI was used to import data directly from the database.
3. **Creation of Visualizations** Different types of visualizations were created in Power BI to represent key metrics:
    - **Pie Chart for Hires by Technology:** Shows the proportion of hires according to different technologies.
    - **Bar Chart for Hires by Year:** Represents the number of hires made each year.
    - **Bar Chart for Hires by Experience Level:** Visualizes hires classified by experience level.
    - **Line Chart for Hires by Country Over Time:** Shows the evolution of the number of hires in selected countries (USA, Brazil, Colombia, Ecuador) over the years.

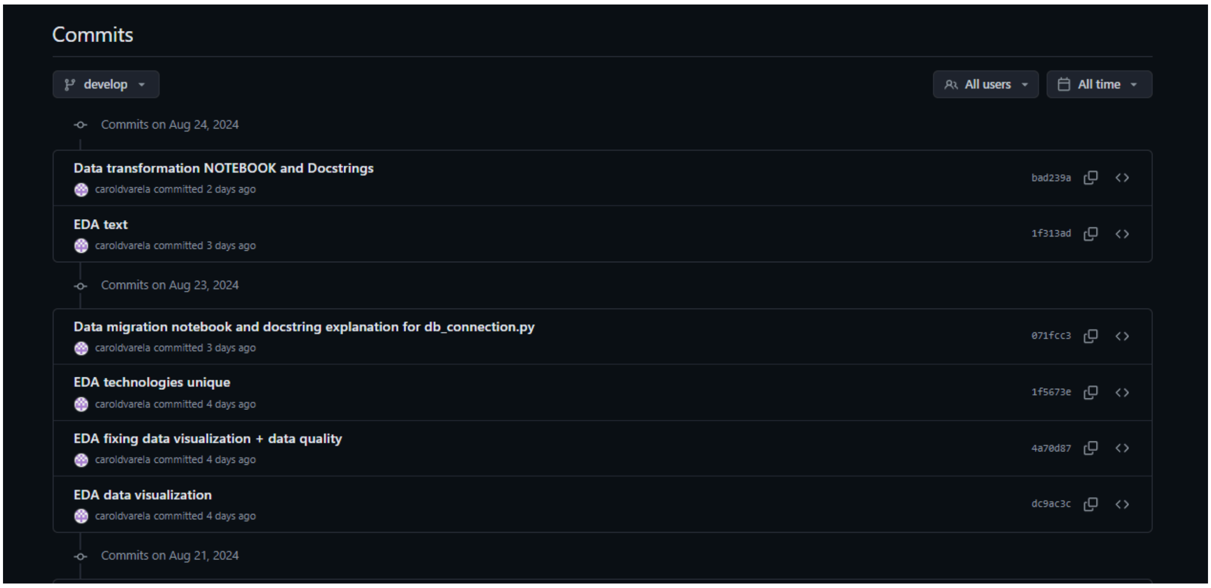


# 8. Conclusion

The workshop was successful in migrating and transforming data from a CSV file to a PostgreSQL database and in developing an interactive dashboard in Power BI. Data migration was smoothly executed, with table creation and insertion of transformed data into the database. Transformations included column normalization, addition of unique identifiers, and technology categorization, facilitating more structured and effective analysis.

The Power BI dashboard has been a key tool for visualizing data in an understandable and dynamic manner. The visualizations created, including pie charts, bar charts, and line charts, provided a clear view of hires by technology, year, experience level, and country. This approach enabled a detailed evaluation of important metrics and facilitated informed decision-making.

In summary, the workshop not only optimized data handling and analysis but also demonstrated the usefulness of interactive visualizations in Power BI for enhancing the understanding of key metrics. The experience highlighted the importance of rigorous data transformation and the effectiveness of visualization tools for strategic analysis and decision-making.



Commits on Aug 24, 2024

**Data transformation NOTEBOOK and Docstrings** — caroldvarela committed 2 days ago — bad239a

**EDA text** — caroldvarela committed 3 days ago — 1f313ad

Commits on Aug 23, 2024

**Data migration notebook and docstring explanation for db_connection.py** — caroldvarela committed 3 days ago — 071fcc3

**EDA technologies unique** — caroldvarela committed 4 days ago — 1f5673e

**EDA fixing data visualization + data quality** — caroldvarela committed 4 days ago — 4a70d87

**EDA data visualization** — caroldvarela committed 4 days ago — dc9ac3c

Commits on Aug 21, 2024