ETL workshop 1

Juan Carlos Quintero (2225339)

Quinto semestre de ingeniería de datos e IA

Javier Alejandro Vergara

ETL (Extract, transform and load)

Universidad Autónoma de Occidente

Santiago de Cali

2024

Introducción

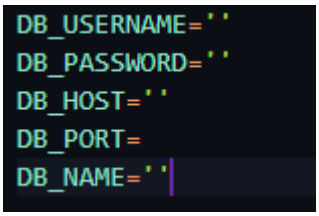
Este documento describe el proceso seguido para resolver el reto del "Python Data Engineer Workshop", diseñado para simular un escenario de entrevista laboral real. El objetivo principal del ejercicio era demostrar habilidades en la extracción, transformación y carga de datos (ETL), así como en la visualización de datos, comenzando con un archivo CSV que contenía información generada de manera aleatoria sobre candidatos que participaron en procesos de selección.

Metodología

La solución al workshop involucró una serie de pasos importantes, desde la preparación del entorno de desarrollo hasta la migración de los datos y la creación de visualizaciones. A continuación, se detallan los pasos principales:

1. Preparación del Entorno de Desarrollo: Se utilizó Jupyter Notebook como el entorno de desarrollo principal, dada su capacidad para integrar código, visualizaciones y explicaciones en un solo lugar de manera interactiva. El lenguaje de programación utilizado fue Python. Para gestionar las librerías y sus dependencias de manera eficiente, se creó un entorno virtual utilizando virtualenv, lo cual permitió trabajar en un entorno controlado y asegurar la consistencia en el proyecto.

2. Gestión de Credenciales para la Base de Datos: Para proteger la información sensible como las credenciales de la base de datos, se emplearon archivos .env con la ayuda de la librería dotenv. Este método permite almacenar de forma segura detalles como el nombre de usuario, contraseña, host, puerto y nombre de la base de datos, evitando que estos datos se expongan directamente en el código.



```
db_username = os.getenv("DB_USERNAME")
db_password = os.getenv("DB_PASSWORD")
db_host = os.getenv("DB_HOST")
db_port = os.getenv("DB_PORT")
db_name = os.getenv("DB_NAME")
```

3. Configuración de la Base de Datos: Se eligió PostgreSQL como sistema de gestión de bases de datos por su capacidad para manejar grandes volúmenes de datos de manera eficiente.

4. EDA: Durante el proceso de análisis exploratorio de datos (EDA), se realizaron varias acciones para comprender mejor las características y distribución de los datos relacionados con los candidatos contratados. A continuación, se detallan los pasos principales y los hallazgos más relevantes.

Carga y Filtrado de los Datos:

El primer paso consistió en cargar los datos desde un archivo CSV y aplicar un filtro para seleccionar únicamente a los candidatos que cumplieran con los requisitos para ser considerados contratados (ambas puntuaciones, tanto en el desafío de código como en la entrevista técnica, superiores o iguales a 7). Este proceso permitió depurar el conjunto de datos y enfocarse en los candidatos más aptos.

```
import pandas as pd

df = pd.read_csv("raw_data/candidates.csv", delimiter=';',encoding='unicode_escape')
df.head()
```

	First Name	Last Name	Email	Application Date	Country	YOE	Seniority	Technology	Code Challenge Score	Technical Interview Score
0	Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern	Data Engineer	3	3
1	Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern	Data Engineer	2	10
2	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus	4	Mid-Level	Client Success	10	9
3	Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea	25	Trainee	QA Manual	7	1
4	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-Level	Social Media Community Management	9	7

We are only interested in candidates who have been hired (both scores greater than or equal to 7), so we will create a data frame with candidates who meet these criteria.

```
df_contracted = df[(df['Code Challenge Score'] >= 7) & (df['Technical Interview Score'] >= 7)].copy()
```

Conversión de Tipos de Datos y Creación de Nuevas Columnas

Se procedió a convertir la fecha de aplicación en un formato de fecha y a extraer el año de aplicación, creando una nueva columna para facilitar análisis posteriores.

```
df_contracted['Application Date'] = pd.to_datetime(df_contracted['Application Date'], errors='coerce')

df_contracted['Application Year'] = df_contracted['Application Date'].dt.year

df_contracted.head()
```

	First Name	Last Name	Email	Application Date	Country	YOE	Seniority	Technology	Code Challenge Score	Technical Interview Score	Application Year
2	Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus	4	Mid-Level	Client Success	10	9	2020
4	Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-Level	Social Media Community Management	9	7	2020
8	Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy	18	Lead	Social Media Community Management	7	10	2018
13	Hilda	Rodriguez	jordan.hyatt@hotmail.com	2020-05-09	El Salvador	16	Junior	System Administration	7	8	2020
22	Crawford	Ullrich	bruce.koch7@yahoo.com	2021-01-09	Dominica	14	Junior	Game Development	8	8	2021

Análisis de Valores Nulos y Resumen Estadístico

Se verificó la existencia de valores nulos y se generó un resumen estadístico de las columnas numéricas para entender mejor la distribución de los datos.

```
df_contracted.isnull().sum()
```

```
First Name                 0
Last Name                  0
Email                      0
Application Date           0
Country                    0
YOE                        0
Seniority                  0
Technology                 0
Code Challenge Score       0
Technical Interview Score  0
Application Year           0
dtype: int64
```

Using the information provided by the functions df.info() and df.isnull().sum(), we can conclude that the dataset has no null values and contains two types of data: objects and integers.

```
numeric_summary = df_contracted.describe()
print(numeric_summary)
```

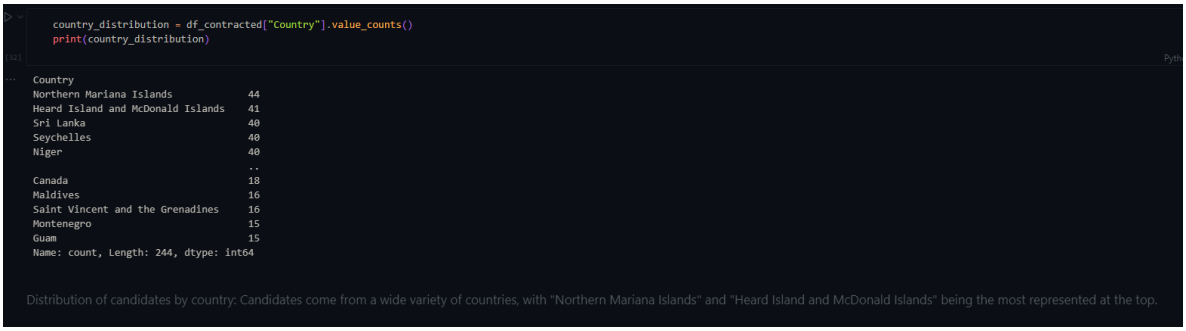
```
                      Application Date          YOE  Code Challenge Score  \
count                             6698  6698.000000           6698.000000
mean   2020-04-10 23:23:40.005972224    15.291281              8.500000
min              2018-01-01 00:00:00     0.000000              7.000000
25%              2019-03-07 00:00:00     8.000000              8.000000
50%              2020-04-09 00:00:00    15.000000              8.000000
75%              2021-05-26 00:00:00    23.000000              9.000000
max              2022-07-04 00:00:00    30.000000             10.000000
std                              NaN     8.843949              1.110748

       Technical Interview Score  Application Year
count                6698.000000       6698.000000
mean                    8.479248       2019.810839
min                     7.000000       2018.000000
25%                     7.000000       2019.000000
50%                     8.000000       2020.000000
75%                     9.000000       2021.000000
max                    10.000000       2022.000000
std                     1.126308          1.315268
```

Summary statistics: Years of experience (YOE): show that candidates have between 0 and 30 years of experience, with a mean of approximately 15 years.

Code challenge and technical interview scores: Both scores range from 0 to 10, with means close to 5, suggesting an even distribution.

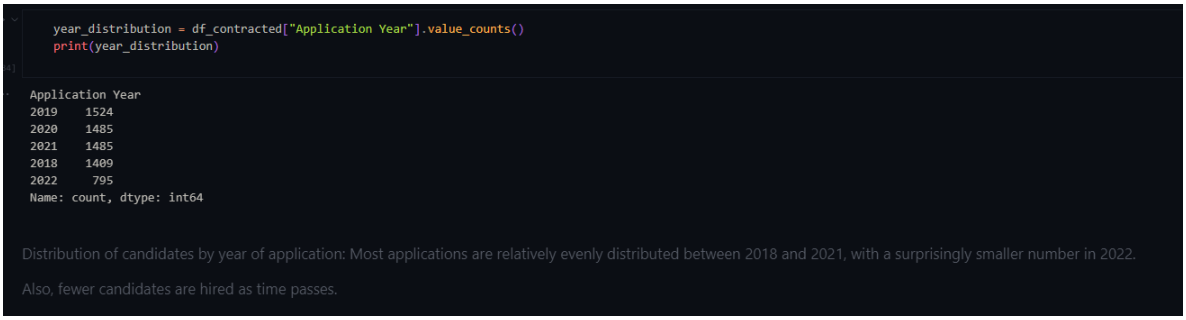
Distribución de los Candidatos por País y Tecnología

Se analizó la distribución de los candidatos contratados según su país de origen y la tecnología en la que se especializan. Esto permitió identificar las áreas tecnológicas más demandadas y los países con mayor representación entre los candidatos.



Distribución de los Candidatos por Año de Aplicación

Finalmente, se analizó la distribución de las contrataciones a lo largo de los años, observando cómo han variado las tendencias de contratación en el tiempo.



5. Migración de Datos: La migración de los datos del archivo CSV limpio a la base de datos en la nube se realizó mediante un script en Python que empleó pandas para la lectura de los datos y SQLAlchemy para conectarse a la base de datos y realizar la inserción de los datos.

6. Análisis y Visualización de Datos: Con los datos ya en la base de datos PostgreSQL, se procedió a realizar el análisis y crear las visualizaciones solicitadas utilizando Power BI. Las visualizaciones generadas incluyen:

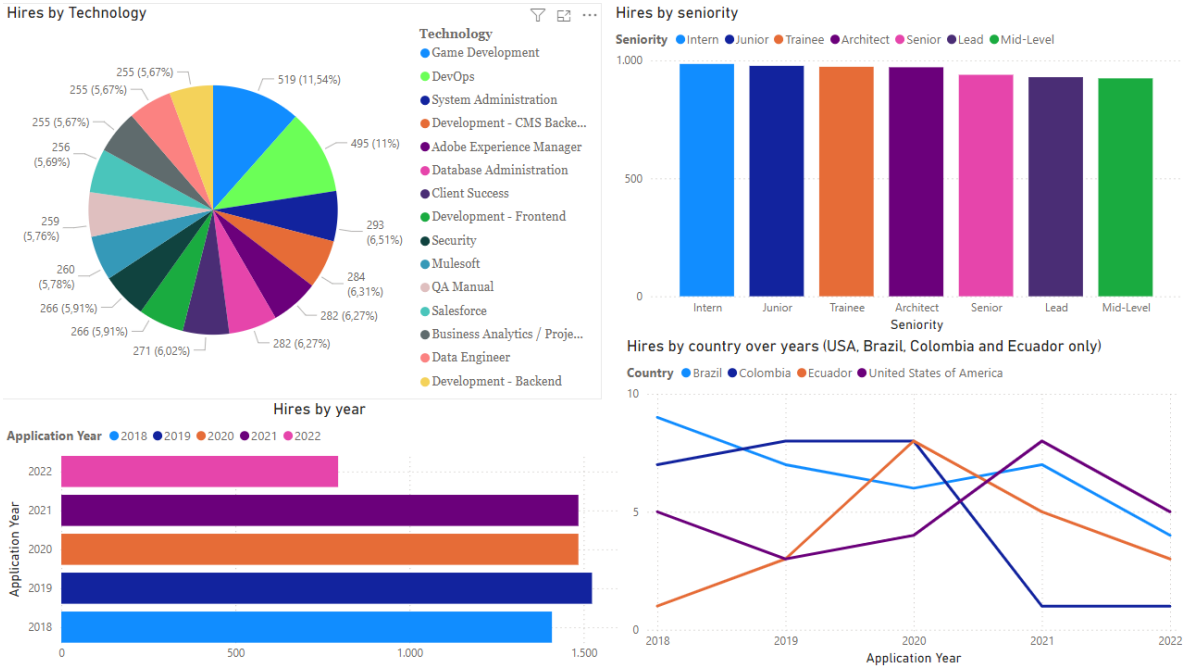
- Distribución de contrataciones por tecnología (Gráfico de tarta)

- Contrataciones por año (Gráfico de barras horizontal)

- Contrataciones por nivel de seniority (Gráfico de barras)

- Contrataciones por país a lo largo del tiempo (Gráfico de líneas múltiples)

Aquí puedes ver el dashboard en detalle:



7. Conclusiones del Análisis: El análisis de los datos de contratación ha permitido obtener una visión detallada de las tendencias de contratación en términos de tecnología, seniority, y distribución geográfica. Se observó que las tecnologías más demandadas fueron Game Development y DevOps, con una distribución uniforme entre diferentes niveles de experiencia. Aunque la contratación se mantuvo constante entre 2018 y 2021, se observó una ligera

disminución en 2022. Geográficamente, Estados Unidos mostró estabilidad, mientras que Brasil y Colombia presentaron una tendencia decreciente y Ecuador experimentó un crecimiento.

Este análisis proporciona una base sólida para la toma de decisiones estratégicas en la gestión de talento y el desarrollo de planes de reclutamiento, así como una orientación clara para futuros análisis y modelos predictivos.