WORKSHOP1
ETL
PYTHON DATA ENGINEER CODE CHALLENGE

Dayanna Vanessa Suarez Mazuera - 2221224

Ingeniería de Datos e Inteligencia Artificial

Semestre 5
Corte 1

ETL
JAVIER ALEJANDRO VERGARA ZORRILLA

Santiago de Cali, Valle del Cauca
2024 - 01

Code challenge for Python Data Engineer

The purpose of this challenge is to do some analysis and manipulations with data from candidates who participated in selection processes for a job and depending if the score obtained in the technical interview and the code challenge is greater than or equal to 7, this candidate could be hired or not hired (this data was randomly generated), and for this ETL (Extract, Transform and Load) and an EDA (Exploration Data Analysis) were realized.

Tasks performed.

1. Extract, Transform and Load (ETL): I performed data extraction from a csv file, and then transformed it into a table and loaded it from postgresql database.
2. Exploratory Data Analysis (EDA): I conducted an exploratory data analysis to better understand the feature distributions and to identify any outliers or missing values.
3. Saving the dataset: I saved the dataset used in the EDA to another table in of the postgresql for future use.
4. Visualization: I created a PowerBI dashboard to analyze the transformed data for a better understanding.

Technologies used.

1. Python: The language used for the workshop.
2. Jupyter notebook: The notebook platform used to make the EDA and other transformations.
3. Visual Studio Code: The chosen code editor for the workshop management and development.
4. PostgreSQL: The database management system used for storing the candidate's data.
5. Power BI: The visualization platform used to make the dashboard.

Architecture.

workshop1

```
├── .gitignore ←Rules for ignore latest versions files
├── README.md ← Principal documentation for the project.
├── config ← Configuration for the connection to the database
│   ├── config.py ←Module with configurate functions
│   └── database.ini ←Configuration file for the connection to the database
├── data ← Output and Input data of the project.
│   ├── candidates.csv ← Candidates raw data.
│   └── candidates_EDA.csv ← Candidates data transformed.
├── documents ← Additional documents of the project.
│   └── document.docx ← Pdf document of the project.
├── notebooks ←Jupyter Notebooks for the Exploratory Data Analysis.
│   └── eda_candidates.ipynb ← Notebook used for the EDA and some transformation.
├── src ← Main code of the project.
│   └── main.py ← Principal script for the process of ETL.
└── visualization ← Visualization files.
    └── workshop_visualization.pbix ← PowerBI dashboard file.
```

Data information.

In this candidates.csv file have 50k rows of data about candidates. The fields we will use are:

- First Name
- Last Name
- Email
- Country
- Application Date
- Yoe (years of experience)
- Seniority

• Technology

• Code Challenge Score

• Technical Interview

Implementation.

1. Start running the main.py file into the src folder (remember to have the database.init file with your database connection information). This file takes care of creating of the database, connecting to it, and the creating of the "candidates" and "candidates_hired" tables, as well as the logic for adding a column called "hired."
2. Run the notebook eda candidates.ipynb, that which contains the exploratory data analysis, which gives a description of what the data is, how it is structured, and some graphs to visualize what the data can tell us. Additionally, agree a column called category_of_technology that contains the categorization of the technologies.

Exploratory Data Analysis (EDA)

Taken the table candidates_hired we can see some information:

General information of the data:

#	Column	Non-Null Count	Dtype
0	Id	50000 non-null	int64
1	firstname	50000 non-null	object
2	lastname	50000 non-null	object
3	email	50000 non-null	object
4	applicationdate	50000 non-null	datetime64[ns]
5	country	50000 non-null	object
6	yoe	50000 non-null	int64
7	seniority	50000 non-null	object
8	technology	50000 non-null	object
9	codechallengescore	50000 non-null	int64
10	technicalinterviewscore	50000 non-null	int64
11	hired	50000 non-null	bool

dtypes: bool(1), datetime64[ns](1), int64(4), object(6)

Columns of the data: 'id', 'firstname', 'lastname', 'email', 'applicationdate', 'country', 'yoe', 'seniority''technology', 'codechallengescore', 'technicalinterviewscore', 'hired'.

Data types:

id	int64
firstname	object
lastname	object
email	object
applicationdate	datetime64[ns]
country	object
yoe	int64
seniority	object
technology	object
codechallengescore	int64
technicalinterviewscore	int64
hired	bool

Graphics

Pie Chart for Technology Values