Workshop -001: Extract

Introduction

This workshop is an exercise of a initial extract process

A few things to consider:

- I ask that you complete this challenge within the timeframe agreed on our conversation.
- **You cannot use tools such as Copilot, Tabnine, Captain Stack, GPT-Code-Clippy, chatGPT, or similar to simplify or generate code to support the challenge. Doing this will be grounds for automatic disqualification.**

Getting Started

Hey, welcome to the **Python Data Engineer** code challenge. In this challenge, I am interested in seeing your knowledge about an extraction process. I will give you some data, and your final objective is to do and extraction process and some transformations.

You will receive a CSV file with data from candidates who participated in selection processes (these data were randomly generated), and you will have to do some analysis and manipulations on top of this data.

You can start coding from scratch, and the technologies we expect to evaluate are described in the technologies section.

What is Expected

I expect that you get the CSV file and create an application to migrate the data to a relational database. Also, you will read the data once stored in the database, create some transformations and store it again in a new table; remember, the data should be stored in a database and all must be done using python.

Steps:

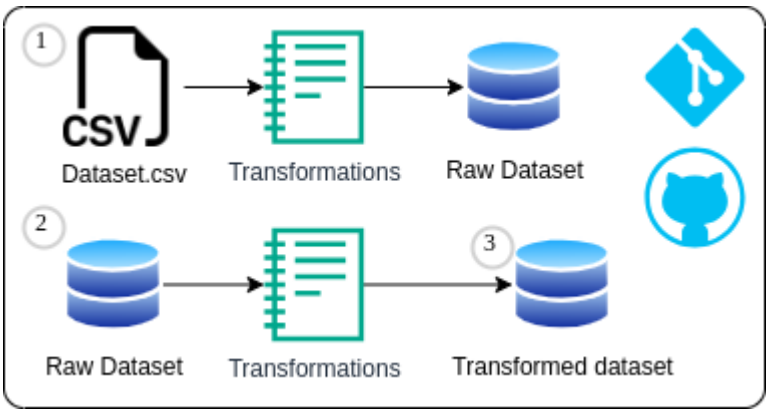
1. Read the csv file using python
2. load the data into the database (staging area)
3. read the table you just created with python
4. do the transformations named bellow
5. load the transformed data into a new table

Technologies

We expect you to use in this challenge:

- Python
- Jupiter Notebook
- Database (you choose)

Diagram



Data

I 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

Transformations:

Create a column hired. consider a candidate HIRED when he has both scores greater than or equal to 7; you should apply this logic to get the correct information. How you will handle this data is on you.

And please remember, all the data here is totally random; we used a public library to generate random information.

Data example

candidates.csv

First Name	Last Name	Email	Application Date	Country	YOE	Seniority	Technology	Code Challenge Score	Technical Interview Score
Bernadette	Langworth	leonard91@yahoo.com	2021-02-26	Norway	2	Intern	Data Engineer	3	3
Camryn	Reynolds	zelda56@hotmail.com	2021-09-09	Panama	10	Intern	Data Engineer	2	10
Larue	Spinka	okey_schultz41@gmail.com	2020-04-14	Belarus	4	Mid-Level	Client Success	10	9
Arch	Spinka	elvera_kulas@yahoo.com	2020-10-01	Eritrea	25	Trainee	QA Manual	7	1
Larue	Altenwerth	minnie.gislason@gmail.com	2020-05-20	Myanmar	13	Mid-Level	Social Media Community Management	9	7
Alec	Abbott	juanita_hansen@gmail.com	2019-08-17	Zimbabwe	8	Junior	Adobe Experience Manager	2	9
Allison	Jacobs	alba_rolfson27@yahoo.com	2018-05-18	Wallis and Futuna	19	Trainee	Sales	2	9
Nya	Skiles	madisen.zulauf@gmail.com	2021-12-09	Myanmar	1	Lead	Mulesoft	2	5
Mose	Lakin	dale_murazik@hotmail.com	2018-03-13	Italy	18	Lead	Social Media Community Management	7	10
Terrance	Zieme	dustin31@hotmail.com	2022-04-08	Timor-Leste	25	Lead	DevOps	2	0
Aiyana	Goodwin	vallie.damore@yahoo.com	2019-09-22	Armenia	24	Intern	Development - CMS Backend	4	9