**FACULTY OF ENGINEERING AND BASIC SCIENCES**
**ACADEMIC PROGRAM: DATA ENGINEERING AND ARTIFICIAL INTELLIGENCE**

**COURSE: ETL (G01)**
**Workshop-1: Data Engineer**

## ✅ Introduction

This workshop simulates a **real job interview code challenge**. It will help you understand what companies expect in recruitment processes and allow you to create a **portfolio project** to showcase on GitHub for your future career.

Your task is to design and implement an **end-to-end ETL process**: extract data from a CSV file, transform it into a **dimensional data model (DDM)**, load it into a **Data Warehouse (DW)**, and finally build reports with KPIs and visualizations that query the DW (not the CSV).

## ✅ Getting Started

Welcome to the Python Data Engineer Challenge.

You will receive a CSV file with 50,000 rows of candidate data from selection processes (randomly generated). Your goal is to:

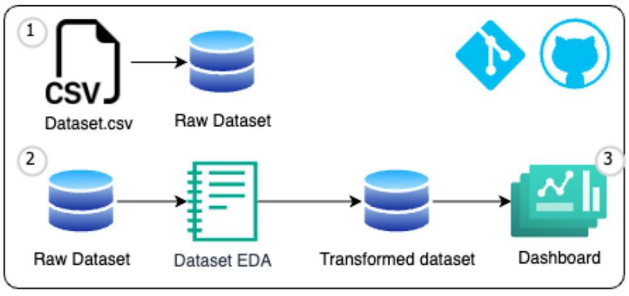
- Design a **Dimensional Data Model (Star Schema)**.
- Load the transformed data into a Data Warehouse (DW).
- Build queries, KPIs, and visualizations directly from the DW.

## ✅ Technologies

You should use:

- Python
- Jupyter Notebook
- Data Warehouse (you choose)

## ✅ Diagram



## 📁 Data Description

You will receive a CSV file containing 50k rows with the following fields:

- First Name
- Last Name
- Email
- Country
- Application Date
- Yoe (years of experience)
- Seniority
- Technology
- *Code Challenge Score*
- *Technical Interview Score*

A candidate is considered **HIRED** if **both scores are ≥ 7**.
You should apply this logic to get the correct information. How you will handle this data is on you.

## 📁 Data Example

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

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 |

## ✅ Task Breakdown

### 1. Dimensional Data Model (DDM)

Design a **Star Schema**:

- **Fact Table**.
- **Dimension Tables**.
- Provide the **diagram + justification** of your design.

### 1. ETL Process

- **Extract:** Load the CSV file in Python.
- **Transform:**
  - Apply the "HIRED" rule.
  - Optional: create tables for dimensions and facts to simplify the Load process.
- **Load:** Insert the transformed data into a DW.

### 3. KPIs & Visualizations

Your reports **must come from the DW** (not the CSV). Expected:

- Hires by technology.
- Hires by year.
- Hires by seniority.
- Hires by country over years (focus on USA, Brazil, Colombia, Ecuador).
- **+2 additional KPIs of your choice** (e.g., % hire rate, average scores, hires by experience range, etc.).

Choose any visualization type, but it must clearly show the information.

### 4. Deliverables in GitHub

Your repository should include:

- **ETL Notebook** (code for Extract, Transform, Load).
- **SQL Queries** to extract KPIs from DW.
- **Star Schema Diagram** (image + explanation).
- **Visualizations** (in notebook or exported as screenshots).
- **README.md** explaining your project, setup instructions, and key decisions.
- **.gitignore** file.

### ✅ Evaluation Criteria

| Item | Description | Weight |
|---|---|---|
| **GitHub Repo Setup** | Repo created, organized, clean structure | 0.3 |
| **Readme** | Clear documentation of approach, setup, usage | 0.3 |
| **Gitignore** | Properly ignores unnecessary files | 0.2 |
| **Dimensional Data Model** | Star schema diagram + justification | 1.0 |
| **Migration to DW** | Correct loading of transformed data into DW | 0.4 |
| **Extracting from DW** | Queries pull data from DW, not CSV | 0.4 |
| **KPIs & Visualizations** | Correct metrics, accurate, clear charts | 1.0 |
| **Documentation** | Explanation of ETL pipeline, challenges, assumptions | 0.4 |
| **Presentation** | Clarity and Structure, Communication and Professionalism | 1.0 |