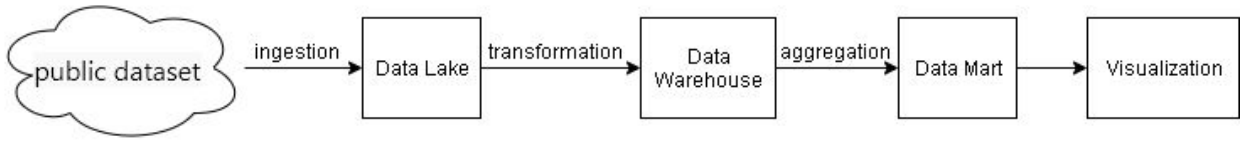
Background

This task is created for data engineer - beginner level. The goal of this task is to measure the understanding of participants about data warehouse and data transformation. Participant will be given the public datasets as a data source and begin to transform and denormalize the data. Finally, the participant will create a data mart which will be visualized due to get some insight.

Scope:

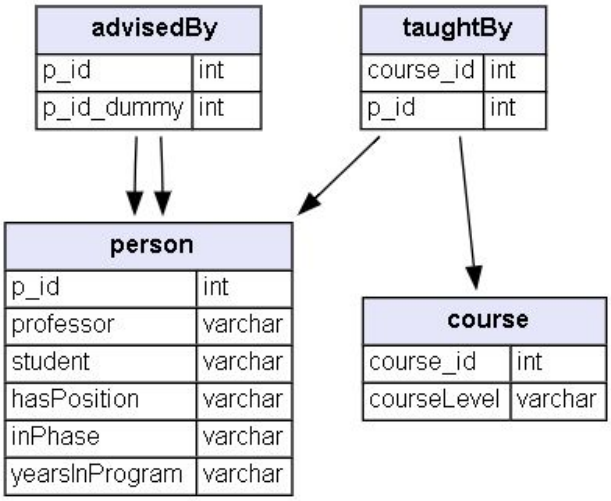


1. Data Ingestion
   Consume public datasets and discover its data.
2. Data Transformation
   - Data Lake
     Store all data which already ingested to specific query-able file format.
   - Transformation (Denormalize Data)
     Joined all data into full and complete dataset
   - Aggregation (Data Mart)
     Created final dataset which ready to be analyzed.
3. Data Visualization
   Visualize data mart to make it easy to get an insight.

Use Case:

UW-CSE Dataset:
This dataset lists facts about the Department of Computer Science and Engineering at the University of Washington (UW-CSE), such as entities (e.g., Student, Professor) and their relationships (i.e. AdvisedBy, Publication).

Syarif Hidayatullah



The datasets are publicly available directly from MariaDB database.
Candidates can use the following credentials:
hostname: relational.fit.cvut.cz
port: 3306
username: guest
password: relational

Another details can be found here: https://relational.fit.cvut.cz/dataset/UW-CSE

Steps:

1. Connect to the public datasets.
2. Consume all datasets and dump into a specific file format csv)
3. Denormalized the data (join) into single completed dataset.
   * can use any programming language (python prefered)
4. Dump the denormalized result into csv file.

Notes

1. Push all your code and the result (CSV) into single zip file.
2. Mandatory to use git. Make sure to use proper git commit comment and commit every major change in your code. Please also exclude unused file using gitignore.
3. Nice to have to use Docker. Create a dockerfile and the requirements library and make sure the jury can run your code inside a docker easily.
4. List all library (if any) into requirements.txt in the root path of your repository.

Syarif Hidayatullah