Project Assignment - Part 1

Roberto Pellungrini, Anna Monreale

October 11, 2022

Introduction

In **Part 1** of the project you are required to create and populate a database starting from .csv files and perform different operations on it. In the following you can find a set of incremental assignments, each one with a brief description of what you are required to produce and what tools you can use for the task.

Build the datawarehouse

answers_full.csv contains the main body of data: a table with data about answers given by students to various multiple-choice questions. In the same table, there are several data regarding the questions, the students and the subject of the questions.

The file **subject_metadata.csv** contains informations about the subject of each question. The subject is given by a list of integers in the main data that can be used to index the **answers_full.csv** to retrieve the topic of the question.

You will have to split and integrate the main file to reproduce the schema in Figure 1.

The goal of the following assignments is to build the schema and deploy it on server lds.di.unipi.it. Beware that, just as in real-life scenario, files may contain missing values and/or useless data.

Assignment 0

Create the database schema in Figure 1 using SQL Server Management Studio in server lds.unipi.it. The name of the database must be *GroupIDHWMart* (example: Group01HWMart).

Assignment 1

Write a python program that splits the content of **answers_full.csv** into the six separate tables: answers, organization, date, subject, user and geography. You will also have to write several functions to perform integration of the main data body. In particular:

- You will have to generate some missing ids, like organizationid and geoid. Use the data that you have available in a suitable way to infer or generate these ids.
- the **iscorrect** attribute is the main measure of the datawarehouse. You can compute its values by comparing the variables answer_value and correct_answer
- the description in the subject table should be a string describing the various topics of the question in subject level order (explore the **subject_metadata.csv** to learn more about that)
- find a way of integrating the continent into the Geography table. You can retrieve the information somewhere, or find a way of providing it yourself
- the Data table should accommodate for both dates of birth of users and for dates of answers. You can clip dates to the day, discarding hours and minutes.

All the above operations must be done WITHOUT using the pandas library.

Assignment 2

Write a Python program that populates the database *GroupIDHWMart* with all teh data you prepared in Assignment 1, establishing schema relations as appropriate.

When you want to deliver your first project, compress the folder and create a single .zip file, named LDS_GroupID.zip. Then send an email to both teachers, with the subject: LDS PART1 Group_Id.

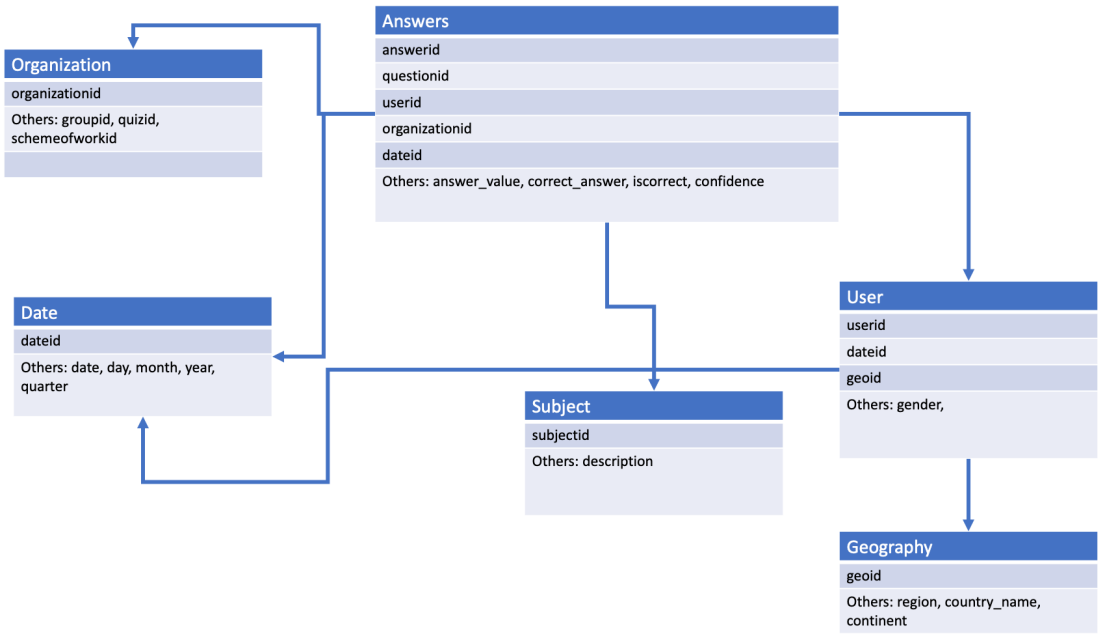

Figure 1: Datawarehouse schema of reference.