Coursework ReportSQL & Data Processing

short line

Kseniya Veka  
ESDE/EHU, 2024

# Introduction

1. Task Description
2. Project Description
3. Tools used
4. Workflow

# 

# Task Description

# 

# Project Description

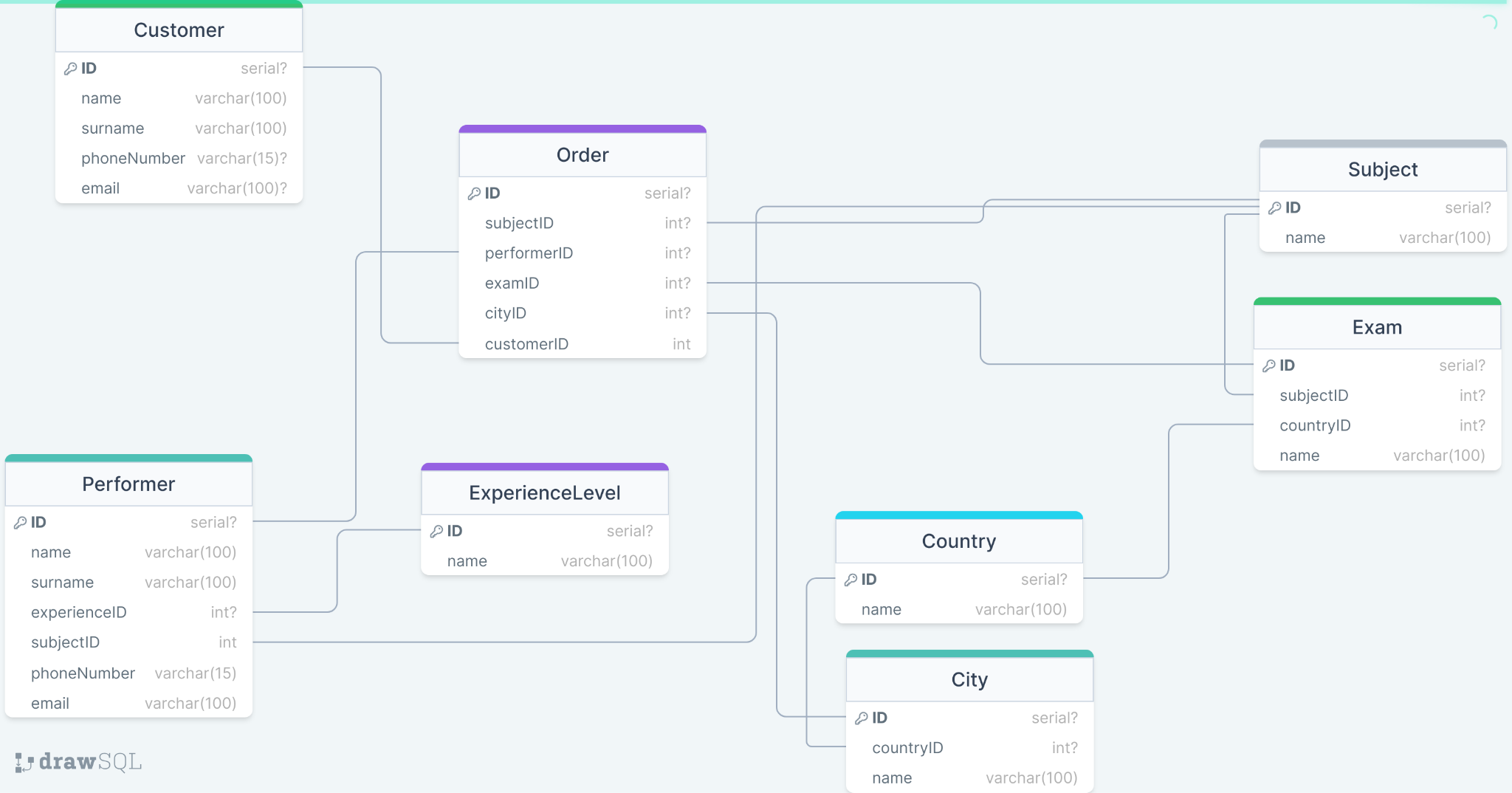
*My application* [*@repet.hub*](https://apple-booklet-260.notion.site/repet-hub-677bfcdbbe8e43c58f4860865c497915) *provides educational services (design, materials creating, video editing, curators' services, SMM) for teachers and online schools.  
  
The business concept suggests that customers (teachers/online schools’ representatives) are going to place an order to receive required services connected with the subject and exam. Customers also have the opportunity to choose the performer according to its specialization and experience.*

*When placing an order, the customer chooses the city of its registration – the company may have numerous offices in different countries and cities.*

*Generally, entities City and Country are only used for legal purposes due to remote working processes (Customer and Performer are not location-dependent).*

My OLTP solution (DB) refers to such entities as:

* Customer
* Performer
* Experience Level
* Subject
* Exam
* Country
* City
* Order



# Workflow

1. **Developing** [**ER diagram**](https://github.com/vekaonelove/SQL_coursework/blob/main/ER%20diagram.jpg)

For this step I have used [drawsql.app](https://drawsql.app/teams/my-team-1313/diagrams/repet), it allows creating entities with attributes of corresponding types.

1. **Creating DB using PostgreSQL -** [**OLTP solution**](https://github.com/vekaonelove/SQL_coursework/blob/main/OLTP%20solution)

As for this step, [drawsql.app](https://drawsql.app/teams/my-team-1313/diagrams/repet) provides an opportunity to export the visual diagram into SQL script: PostgreSQL and MySQL conversions are well-performed automatically

1. **Creating datasets: 2 .csv files containing data according to DB entities:** [**file1**](https://github.com/vekaonelove/SQL_coursework/blob/main/file1.csv) **and** [**file2**](https://github.com/vekaonelove/SQL_coursework/blob/main/file2.csv)

While generating 2 datasets in .csv format I have used [mockaroo.com](http://mockaroo.com)

First dataset - file1.csv - contains data on Customer, Performer, ExperienceLevel and Subject. It is used to describe Customer-Performer connectivity.

Second dataset contains data on Country, City, Exam, and Order. This one is considered to be administrative.

1. **Writing a *rerunable* ETL script to load data from .csv files to our DB using** [**Spring Batch**](https://github.com/vekaonelove/SQL_coursework/tree/main/ETL1)

I have used Java + Spring to connect to the database, extract data from files, transform it (name the columns properly, add constraints to remove duplicates and put the values into proper columns).

I have also configured entity (model) classes for every table that I had to work with and created their instances before putting data into DB tables.

1. **Building a snowflake DWH based on our initial DB** [**(OLAP solution)**](https://github.com/vekaonelove/SQL_coursework/blob/main/OLAP%20solution)
2. In this case I have marked the Order table as the business fact: it contains all the information that could be valuable for business purposes.
3. **Writing a second ETL script to move data from OLTP DB to the OLAP one using Spring Batch**

Also using Java and Spring, thus without Entities, being connected to 2 databases at one time. The script uses several methods, but tends to avoid copy-paste and implements a general solution then applied to every table separately.

1. **Creating a PowerBI report with visualization**

Analyzing Order geography – the amount of orders per country (map), Exams rating according to Country, Exams rating according to Subject, and ExperienceLevel of employees according to each Subject.

1. **Writing detailed documentation on the project and steps performed**

# 