Full Name: Date:

Age:

**Purpose**

This challenge aims to assess how to solve problems. By resolving it, you will show us:

* Its ability to articulate business issues in database queries.
* Its ability to extract data from a source, process it, and save it to a new source.

**About the challenge**

Stack Overflow is a widely known platform in the technology community and allows users to ask questions and also respond to them. In addition, they can, through registration and active participation, vote on more or less useful questions and answers.

You've probably already accessed it to remedy any code doubts you had.

Every year Stack Overflow does a search with its developer community on a variety of topics, ranging from their technological preferences to professional issues. And we're super curious to know what the developers are talking about. We want to know what technologies they use, how they communicate, how much they earn on average, where they live and a few things.

Your challenge is to help us answer these questions using the search results applied in January 2018. We divide the challenge into two main parts:

1. Populate a database from raw search data (we'll already give you the database structure)

2. Perform queries on the database to kill our curiosity

**Assembling the database**

We will give you a text file (CSV format) containing a portion of the search results performed by Stack Overflow and another text file (CSV format) containing the description of the answer columns present in the first file (i.e. it tells you which Questions were asked and that generated the answers).

You will use a programming language to read this file, process it according to the business rules outlined below, and then enter that data into a database of your choice (see the Stack of technologies section).

We'll give you the database relationship entity model, but it will be up to you to assemble the SQL code that implements that model in the bank.

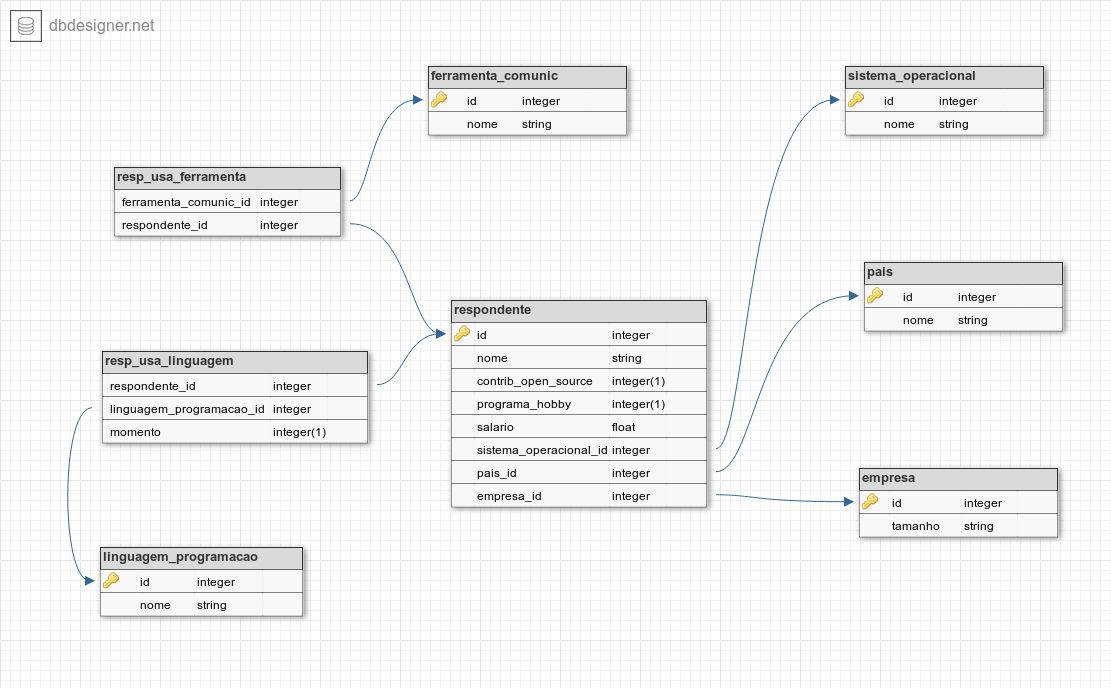
**Data source**

You will find in the attachment of this project two files, the first of which contains a sample of only 10000 lines of answers to the search, and the second, an explanation of the meaning of the response columns.

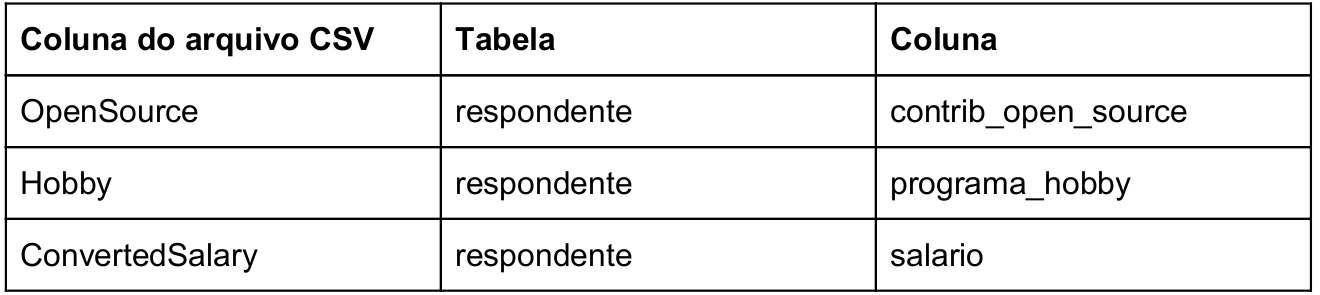
The first file is named *Base\_de\_respostas\_10k\_amostra. csv* and the second, *base\_de\_conhecimento. csv.* If you want to see the full results of the search, simply access this  [kaggle](https://www.kaggle.com/stackoverflow/stack-overflow-2018-developer-survey)link.

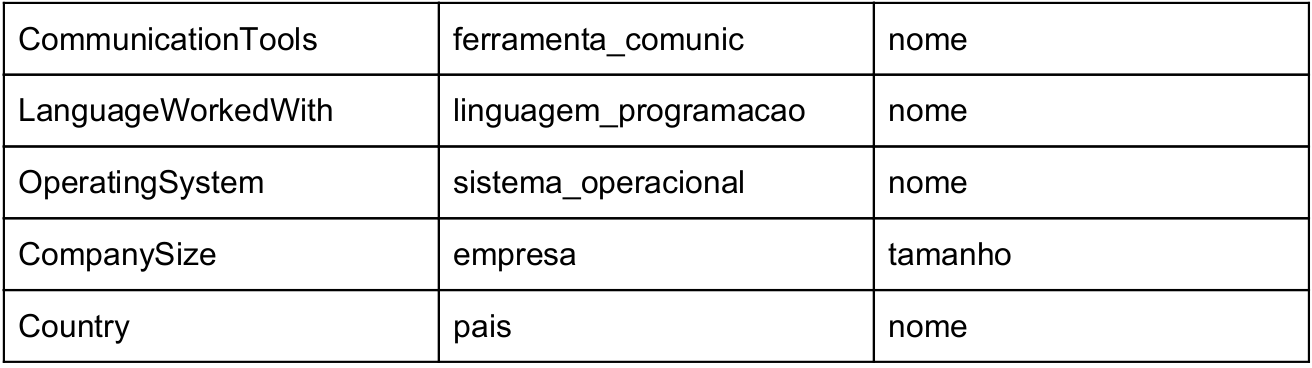
**Database structure**

The image below contains the structure of the database that you will implement. You can also access the larger-sized image in the *MER-summer-job. png* file , which is attached to the project.



The table below maps the fields of the CSV file to the tables in the database. That way, you'll know exactly what to search for and analyze:





**Business rules**

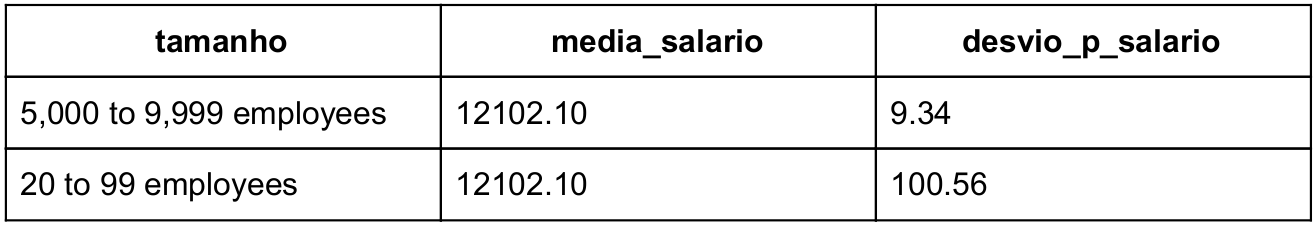
* Empty salary or with "NA" value should be converted to zero (0.0).
* Salary should always be calculated in reais and monthly. For this calculation you will use the column *ConvertedSalary*, which contains the yearly salary. Consider that 1 dollar is equivalent to R $3,81.
* The respondent's name must follow the *responder rule [number]*  . (e.g.: *respondente\_1, respondente\_2*, *respondente\_3*). The criterion for generating that number is all yours.
* Each row in the *Linguagem\_programacao* table must contain a single programming language.
* Each row in the *ferramenta\_comunic* table should contain only one communication tool.

It is important to note that in some response fields there are multiple results, such as in the *Languageworkedwith*column, which contains several programming languages in a row. In these cases, you must break the string at the points that there are semicolons (";").

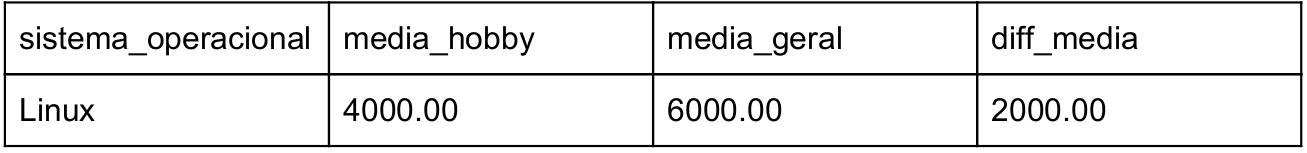
**Questions to**  **be**  **answered**

With your bank structure ready, you can perform SQL queries on the bank you created and kill our curiosity. The list below contains everything we need to know:

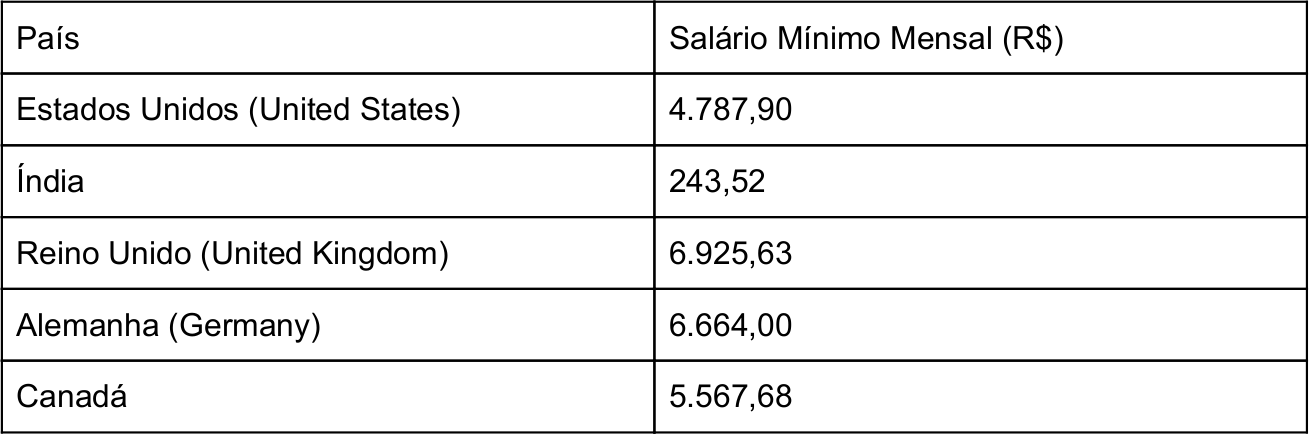
1. How many respondents from each country?
2. How many users who live in "United States" like Windows?
3. What is the average salary of users who live in Israel and like Linux?
4. What is the average and standard deviation of the salary of users who use Slack for each available company size? (Hint: The result should be a table similar to the one shown below)



1. What is the difference between the average salary of Brazilian respondents who think that creating code is a hobby and the average salary of all Brazilian respondents grouped by each operating system they use? (Hint: The result should be a table similar to the one shown below)



1. What are the top 3 technologies most used by developers?
2. What are the top 5 countries in question of salary?
3. The table below contains the monthly minimum wages of five countries present in the data sample. Based on these values, we would like to know how many users earn more than 5 minimum wages in each of these countries.



**Stack of Technologies**

Our team is very diversified in terms of technology. For this project specifically, we selected the following stack of technologies for you to use:

* **Programming language**: Any free language (suggestions: Python, Java, Scala, Ruby)
* **Database:** Any free relational database (e.g. MySQL, PostgreSQL, SQLite, MariaDB, etc.)
* **Database Query Language:** SQL

We know that there is a lot of code available on the Internet and that they often help us solve challenges we face when developing projects. **But, be careful!** We want to know quite the code that you are able to develop, using your analytical ability and creativity. We do not expect a specific code to solve this challenge, but rather that it reflects its knowledge.

Make yourself comfortable if you want to use an IDE to model your database, that is, create your table structure. It's only interesting that you make this clear in your final report.

**What do we expect to see at the end?**

Our team is curious to see your project. We hope your final deliverable contains the following items:

1. An introductory file explaining the overview of your project and which technologies you have used (e.g. PostgreSQL 9.6).

2. A file containing the answers to the questions we made above.

3. The SQL queries you performed on the bank to answer the questions.

4. The files with your code used to read the text files and publish the data in the bank.

If you have difficulty completing your project, we strongly encourage you to send us your progress (SQL queries, code, description of how to solve the problem, etc.).

**COME TO AME! =)**