Homework #4

Prepared by Dmitry Chub [dmchub@yandex.ru](mailto:dmchub@yandex.ru)

* Deadline:
  + Soft (-50 % points) 21st December 2022
  + Strict (-100% points) 23rd December 2022

This task is devoted to work with databases, JDBC and related technologies. All the problems are the parts of one big task, thus you need to complete them sequentially.

## Description of the data

[Link to the data source](https://disk.yandex.ru/d/aC5q2K2UO42a_w)

These tables are filled with the real data collected by Yandex in 2014-2015. This dataset represents information about HTTP requests made by Yandex users. There are two tables of different formats:

1. user\_logs

1. User IP (**STRING**)

2. Query time (**TIMESTAMP**), timestamp format: “**YYYYMMddhhmmss**”

3. HTTP-query (**STRING**)

4. Web page size(**INT**)

5. HTTP status code (**INT**)

6. Information about client’s application and browser (**STRING**)

***Important****: browser info is contained in the starting part of 6th field (from first symbol to the first whitespace character), the following substring has nothing to do with the browser. There are three tab characters between IP and query time.*

2. user\_data

1. User IP (**STRING**)

2. User browser (**STRING**)

3. Gender (**STRING**) ‘male’, ‘female’

4. Age (**INT**)

## Problem A. Getting ready [10 points]

First of all you need to implement the db loading.

1) Create DB and all the tables **[3 points]**

2) Download the data from Yandex Disk **[2 points]**

3) Load tables with downloaded data **[5 points]**

## Problem B. Queries [15 points]

This is the main part of the task. You need to implement the following queries:

| **#** | **Query description** | **Points** |
| --- | --- | --- |
| 1 | Find top-N most active users, N is a parameter of the query | 3 |
| 2 | Find the number of queries by men and women for a given day, day is a parameter of the query | 3 |
| 3 | Find the three most popular websites for users under age A for the last B days, A&B are parameters | 3 |
| 4 | Insert new entry to user\_logs table | 3 |
| 5 | Delete all entries before some day, day is a parameter | 3 |

## Problem C. Data visualization [5 points]

* For each of queries 1-3 your solution should generate an Excel file with query results and save it to the “Reports” folder. Heading needs to have a style different from data cells style **[2 points]**
* For each of queries 1-3 your solution should generate a chart with results and save it to the “Charts” folder **[3 points]**

## Problem D. CI [7 bonus points]

Your task is to set up CI of your project using GitLab CI. This task requires you to use GitLab. However, you can use other platforms (Github for example) and set up CI there with different CI system and complete the same stages.

* For the correct CI-job start use runners with ‘docker-atp’ [tag](https://docs.gitlab.com/ee/ci/yaml/index.html#tags)
* Stages for CI:
  + Validation and project compilation with Maven
  + Downloading the files with data from Yandex Disk (don’t forget to save it for the next stage with Gitlab CI cache)
  + Creation and filling the DB [problem A]
  + Queries + Excel tables and charts building (All the Excel tables and charts from problem C need to be saved as artifacts) [problems B and C]