# Big Data Developer Homework

## Introduction

* To complete the task, we recommend using Scala or Python and Apache Spark. However, you are free to choose any other tool or programming language.
* Completing all the parts of the homework is not mandatory. We strongly encourage you to submit your results even if you did not manage to complete everything.
* To submit your homework, please send back a zip file with an application source code from a Git repository.
* Your application should be deployable and executable from the command line in a Linux environment.
* We will consider not only the correctness of algorithms but also compliance with good coding practices and overall code quality.
* Covering code with unit tests is highly preferred

## Data

* Raw data is in the raw\_data folder.
* Assume that new data files will be uploaded constantly to the raw\_data directory.
* Files are filled with information on impressions (impressions\_processed\_dk\_) and clicks (clicks\_processed\_dk\_\*).
* One record in the files represents one impression/click.
* The file format is parquet <https://parquet.apache.org>

## Task

Create an application that calculates the count of impressions and clicks by date and each hour of the day, for a specific user-agent value „some user agent“. The user agent field is in the device\_settings structure field user\_agent.

* Application can take user agent as an optional parameter and produce results to Apache Kafka that should be running in Docker container.
* Output Kafka topic should contain user-agent as a key.
* datetimes of data inside the files can be determined from a file name: impressions\_processed\_dk\_20220526113212045\_172845633-172845636\_1.parquet  
  , date and time format - „2022-05-26 11:32“.
* If the application is run multiple times with the same user-agent parameter – it should not process already processed files, but only process newly added files since last run.