# Data scientist test task.

Given two files: *train\_data.parquet* and *eval\_data.parquet*.

The first one contains the dataframe with 90 rows and 100 columns: partner\_1, …, partner\_100. Every column of the dataframe represents the number of active customers on a partner for each day for the last 90 days. Some partners are relatively new, so they have 0 customers on some first days (the number of their active users starts to grow from some day different from 0). Some other partners are relatively old, so they already have some customers on day 0. Finally, some other partners don’t have customers on day 0, but start to grow immediately.

**Part 1 of the task.**   
It is required to perform clusterization of the partners based on the given data. Please, provide the code in Python with all necessary steps clearly stated (e.g., data reading, feature preparation, clusterization, visualization, etc.). Report the results in any form and explain all the choices made during the clusterization (e.g., the number of clusters).

**Part 2 of the task.**

Implement the classification of the data in “eval\_data.parquet” as follows. Use the data in train\_data.parquet as the training set and the obtained cluster values at the labels. Then apply some classification algorithm (any of your choice) for the data contained in the file “eval\_data.parquet”. This file contains the dataframe of the same structure and type as previously, but with only 20 columns (partner\_101, …, partner\_120) and only 30 rows (representing the data of the last 30 days). Write the results of the classification in the file “eval\_data\_classification.txt” having two columns: “partnerID” and “class”, where partnerID represents the ID of the partner (partner\_101, …, partner\_120) and class represents the obtained class in the numerical form.

You can implement the code in any format (e.g., Jupyter notebooks or py-files). GUI or testing of the code are not required. For any questions, please do not hesitate to contact me.