First group exercise

In this side group exercise, the objective is to prepare a short business document comparing different approaches processing the Stack Overflow data.

The high-level business case is to get a realistic picture of the online activity, popularity and the overall performance of the examined organizations through monitoring them in different data sources. In the context of Stack Overflow, this is the number of questions related to an organization.

The primary approach to this, which was also part of the weekly core task, was through question tags. But this is only one way to tell if an issue is related to the organization.

This can only work well if the questioner correctly fills in every attribute related to the question, such as tagging. This has a number of positive benefits for the questioner, as it can get the question into communities where questions are answered in relation to a particular technology, or it can be found in contexts where the user would not have targeted the question on its own.

On the other hand, it is possible that the questioner has not bothered to fill in these extra fields, so there may be questions in the platform that, although they belong to the organization, do not appear in the tag method.

The other method is more direct, as it is based on analyzing the question text description itself to see if the name of the organization is mentioned.

For this approach, the first step is to create a new model in the intermediate layer. It is necessary to determine which column contains the question description in the data source and build a model that map the questions to the organizations based on this column.

Once the model is complete, it is necessary to compare the two methods and present the findings in a short document in a form that is also easy for a business user to understand.

For this purpose, in addition to the descriptions, it may be interesting to provide a table or some specific examples to illustrate the differences. It may also be interesting to consider the quantity and the ratio of the two solutions and the possibility of implementing both.

To produce these summaries, it is necessary to write aggregated queries on the tables to support the results. This can be a 1:1 comparison of an organization for the two approaches or aggregated lists.

At the end, it is also interesting to compare the process cost of the two methods. This may be interesting that one approach may return more accurate results, but at a much higher process cost. In this case, it should be considered to choose the less accurate one, but with less process cost.